

500.41226X00



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): MATAUBAYASHI, et al.  
Serial No.: Not assigned  
Filed: February 25, 2002  
Title: METHOD OF SEARCHING SIMILAR DOCUMENT, SYSTEM  
FOR PERFORMING THE SAME AND PROGRAM FOR  
PROCESSING THE SAME  
Group: Not assigned

LETTER CLAIMING RIGHT OF PRIORITY

Honorable Commissioner of  
Patents and Trademarks  
Washington, D.C. 20231

February 25, 2002

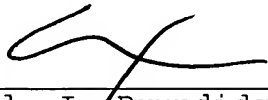
Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the  
applicant(s) hereby claim(s) the right of priority based on  
Japanese Application No.(s) 2001-128934 filed April 26, 2001.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP

  
\_\_\_\_\_  
Carl. I. Brundidge  
Registration No. 29,621

CIB/amr  
Attachment  
(703) 312-6600

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日  
Date of Application:

2001年 4月26日

出 願 番 号  
Application Number:

特願2001-128934

[ST.10/C]:

[JP2001-128934]

出 願 人  
Applicant(s):

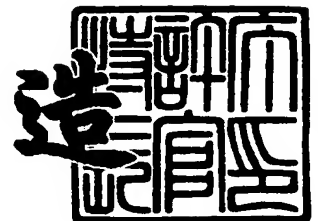
株式会社日立製作所

11046 U.S. PTO  
10/081203  
02/25/02

2002年 1月29日

特 許 庁 長 官  
Commissioner,  
Japan Patent Office

及 川 耕 造



出証番号 出証特2002-3002326

【書類名】 特許願

【整理番号】 K01003091

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/30

【発明者】

【住所又は居所】 神奈川県川崎市幸区鹿島田 890 番地 株式会社日立製作所 ビジネスソリューション事業部内

【氏名】 松林 忠孝

【発明者】

【住所又は居所】 神奈川県川崎市幸区鹿島田 890 番地 株式会社日立製作所 ビジネスソリューション事業部内

【氏名】 多田 勝己

【発明者】

【住所又は居所】 神奈川県川崎市幸区鹿島田 890 番地 株式会社日立製作所 ビジネスソリューション事業部内

【氏名】 里 佳史

【発明者】

【住所又は居所】 神奈川県川崎市幸区鹿島田 890 番地 株式会社日立製作所 ビジネスソリューション事業部内

【氏名】 稲場 靖彦

【発明者】

【住所又は居所】 神奈川県横浜市戸塚区戸塚町 5030 番地 株式会社日立製作所 ソフトウェア事業部内

【氏名】 野田 十悟

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100083552

【弁理士】

【氏名又は名称】 秋田 収喜

【電話番号】 03-3893-6221

【手数料の表示】

【予納台帳番号】 014579

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 類似文書検索方法及びその実施システム並びにその処理プログラム

【特許請求の範囲】

【請求項 1】 指定された文書と類似する文書を検索する類似文書検索方法において、

所望の検索内容を含んだ種文書から特徴語の候補となる特徴語候補を抽出するステップと、前記抽出された特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該種文書の特徴語として抽出するステップと、

前記抽出された種文書の特徴語と登録文書の特徴語との間の類似度を算出するステップと、前記算出された類似度算出結果を検索結果として出力するステップとを有することを特徴とする類似文書検索方法。

【請求項 2】 前記抽出された特徴語候補に対応する特徴語にその構成特徴語を示す構成特徴語情報が登録されている場合に、当該特徴語候補が複合特徴語であると判定することを特徴とする請求項 1 に記載された類似文書検索方法。

【請求項 3】 前記抽出された種文書の構成特徴語に一致する登録文書の特徴語について、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を算出するステップを有し、前記重み係数を乗じた類似度を算出することを特徴とする請求項 1 または請求項 2 のいずれかに記載された類似文書検索方法。

【請求項 4】 指定された文書と類似する文書を検索する類似文書検索システムにおいて、

所望の検索内容を含んだ種文書から特徴語の候補となる特徴語候補を抽出する文書解析処理部と、前記抽出された特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該種文書の特徴語として抽出する特徴語抽出処理部と、

前記抽出された種文書の特徴語と登録文書の特徴語との間の類似度を算出する種文書類似度算出処理部と、前記算出された類似度算出結果を検索結果として出

力する検索結果出力処理部とを備えることを特徴とする類似文書検索システム。

【請求項 5】 前記抽出された特徴語候補に対応する特徴語にその構成特徴語を示す構成特徴語情報が登録されている場合に、当該特徴語候補が複合特徴語であると判定する複合特徴語判定処理部を備えることを特徴とする請求項 4 に記載された類似文書検索システム。

【請求項 6】 前記抽出された種文書の構成特徴語に一致する登録文書の特徴語について、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を算出する重み係数算出処理部を備え、前記重み係数を乗じた類似度を算出することを特徴とする請求項 4 または請求項 5 のいずれかに記載された類似文書検索システム。

【請求項 7】 指定された文書と類似する文書を検索する類似文書検索システムとしてコンピュータを機能させる為のプログラムにおいて、

所望の検索内容を含んだ種文書から特徴語の候補となる特徴語候補を抽出する文書解析処理部と、前記抽出された特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該種文書の特徴語として抽出する特徴語抽出処理部と、

前記抽出された種文書の特徴語と登録文書の特徴語との間の類似度を算出する種文書類似度算出処理部と、前記算出された類似度算出結果を検索結果として出力する検索結果出力処理部としてコンピュータを機能させることを特徴とするプログラム。

【請求項 8】 前記抽出された特徴語候補に対応する特徴語にその構成特徴語を示す構成特徴語情報が登録されている場合に、当該特徴語候補が複合特徴語であると判定する複合特徴語判定処理部としてコンピュータを機能させることを特徴とする請求項 7 に記載されたプログラム。

【請求項 9】 前記抽出された種文書の構成特徴語に一致する登録文書の特徴語について、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を算出する重み係数算出処理部としてコンピュータを機能させることを特徴とする請求項 7 または請求項 8 のいずれかに記載されたプログラム。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は指定された文書と類似する文書を検索する類似文書検索システムに関し、特にユーザから指定された文書に記述されている特徴語を含む文書を類似文書として文書データベースの中から検索する類似文書検索システムに適用して有効な技術に関するものである。

【 0 0 0 2 】

【従来の技術】

近年、組織内での業務の効率化や、業務の質を向上させる為に、組織内の個人の知識を共有し、再利用することを目的とする知識管理システムへの要求が高まってきている。

【 0 0 0 3 】

特に企業内で活用する知識管理システムに対しては、有識者の経験やノウハウ等を文書化し、知識として共有、活用することへの要望が高まっており、組織内で非定型に蓄えられた大量の知識の中から、ユーザが所望するものを簡単にかつ適切に取得する高精度な検索機能が重要になってきている。

【 0 0 0 4 】

このような要求に応える技術として、ユーザが自分の所望する内容を含んだ文書（以下、種文書と呼ぶ）を例示し、その文書と類似する文書を検索する類似文書検索技術が注目されている。

【 0 0 0 5 】

類似文書検索の方法としては、例えば、文書内に出現する単語（以下、特徴語と呼ぶ）の出現頻度を要素とするベクトル（以下、特徴ベクトルと呼ぶ）を用いて文書間の類似度を算出する技術（以下、従来技術 1 と呼ぶ）が、“Information Retrieval” (William B. Frakes, Prentice Hall PTR, pp.363~376) に開示されている。

【 0 0 0 6 】

従来技術 1 の概要は次の通りである。文書データベースに文書を登録する際に、登録対象となる文書中に含まれる特徴語の出現頻度を登録文書の特徴ベクトル

(以下、登録文書特徴ベクトルと呼ぶ)として作成しておく。

【0007】

類似文書の検索時は、検索条件として指定された種文書の特徴ベクトル(以下、種文書特徴ベクトルと呼ぶ)と各登録文書特徴ベクトルとのベクトル空間内においてなす角度の余弦を、文書間の類似度として算出する。

【0008】

図20は従来技術1の処理手順の一例を示す図である。以下、従来技術1の処理手順を図20のPAD(Problem Analysis Diagram)図を用いて説明する。

【0009】

まずステップ200において、文書の登録処理か類似文書の検索処理かを判定する。そして、文書の登録処理と判定された場合には登録文書特徴ベクトル生成ステップ210を実行し、登録対象文書に対する登録文書特徴ベクトルを生成する。

【0010】

また、ステップ200において類似文書の検索処理と判定された場合には、種文書特徴ベクトル生成ステップ220を実行し、検索条件として指定された種文書に対する種文書特徴ベクトルを作成する。

【0011】

次にステップ221を実行し、全登録文書に対して類似度算出ステップ222を繰り返し実行する。類似度算出ステップ222では、前記種文書特徴ベクトルと登録文書特徴ベクトルが、ベクトル空間内においてなす角度の余弦を文書間の類似度として算出する。

【0012】

図21は従来技術1における特徴ベクトル生成処理の一例を示す図である。以下、図20に示した登録文書特徴ベクトル生成ステップ210及び種文書特徴ベクトル生成ステップ220として実行される従来技術1における特徴ベクトル生成処理について、図21に示したPAD図を用いて説明する。

【0013】

特徴ベクトル生成処理では、まずステップ301において、特徴ベクトルの生



成処理対象となる文書を読み込む。次にステップ302において、上記ステップ301で読み込まれた処理対象文書から特徴語を抽出する。

【0014】

そしてステップ303において、上記ステップ302で抽出された各特徴語の出現頻度を計数する。最後にステップ304において、上記ステップ302で抽出された各特徴語と、上記ステップ303で計数した各特徴語の出現頻度を特徴ベクトルの要素として格納する。以上が、従来技術1の処理手順である。

【0015】

図22は従来技術1の概要を示す図である。以下、図22を用いて従来技術1の処理例を説明する。

【0016】

従来技術1では、まず処理要求判定ステップ410において、入力された処理要求が登録処理であるか、或いは検索処理であるかを判定する。そして、入力された処理要求が登録処理である場合には、ステップ210が実行される。

【0017】

ステップ210では、登録用文書1及び文書2中に含まれる特徴語を抽出すると共に各文書内での出現頻度を計数し、各文書に対応する登録文書特徴ベクトル403及び404を生成する。

【0018】

ここで、登録文書特徴ベクトル403“文書1(“LAN”、1) (“構築”、1) …”は、「文書1」の特徴ベクトルであり、特徴語“LAN”が1回、特徴語“構築”が1回出現していることを表している。

【0019】

また、前記処理要求判定ステップ410で類似文書の検索処理と判定された場合には、検索条件で指定された種文書406から特徴語を抽出し、ステップ220で該種文書に対応する種文書特徴ベクトル407を生成する。

【0020】

次に、種文書特徴ベクトル407と前記ステップ210で生成された各登録文書の登録文書特徴ベクトルとのなす角の余弦を類似度として算出する。

【0021】

一般に、2つのベクトルA及びベクトルBのなす角の余弦は、数1の様に算出される。ここで“ $A \cdot B$ ”は、ベクトルAとベクトルBの内積を表し、“ $|A|$ ”は、ベクトルAの大きさを表す。

【0022】

【数1】

数1

ベクトルAとベクトルBのなす角度の余弦

$$= \frac{A \cdot B}{|A| \times |B|}$$

【0023】

図22に示した種文書特徴ベクトル407と登録文書特徴ベクトル403及び登録文書特徴ベクトル404のなす角の余弦は、ベクトルAを種文書特徴ベクトル407、ベクトルBを登録文書特徴ベクトル403または登録文書特徴ベクトル404として、それぞれ数2、数3の様に算出される。

【0024】

【数2】

数2

$$= \frac{1 \times 0 + 1 \times 0 + 1 \times 1 + 1 \times 0}{\sqrt{1^2 + 1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2}} = \frac{1}{2\sqrt{6}} = 0.204$$

【0025】

【数3】

数3

$$= \frac{1 \times 1 + 1 \times 1 + 1 \times 1 + 1 \times 0}{\sqrt{1^2 + 1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2}} = \frac{3}{2\sqrt{5}} = 0.670$$

【 0 0 2 6 】

この結果として、種文書に対する各登録文書の類似度算出結果 4 0 8 が出力される。以上が、従来技術 1 の処理例である。

【 0 0 2 7 】

以上説明した様に従来技術 1 によれば、登録文書中に含まれる特徴語を抽出した登録文書特徴ベクトルを予め作成しておき、検索条件として指定された種文書に対応する種文書特徴ベクトルとの余弦を類似度として算出することで、文書データベース中から内容の類似する文書を検索することができる。

【 0 0 2 8 】

【発明が解決しようとする課題】

しかし従来技術 1 では、特徴ベクトルの要素である特徴語が複数の単語で構成されている場合に、検索漏れが発生するという問題がある。

【 0 0 2 9 】

図 2 3 は従来技術 1 の問題点を示す図である。以下、図 2 3 を用いて、従来技術 1 の問題点を説明する。本図では、文書 3 「地図情報閲覧ソフトを開発、発売した A 社は、・・・」及び文書 4 「多くの地図閲覧ソフトが発売されているが、・・・」が登録された文書データベースに対して、種文書「最新の地図閲覧ソフトについて」が入力された場合の例を表している。

【 0 0 3 0 】

まず文書の登録処理として、ステップ 2 1 0 が実行され、各文書に対応する登録文書特徴ベクトル 4 0 3 a 及び 4 0 4 a が生成される。本図に示した例では、文書 3 に対応する特徴ベクトル 4 0 3 a として“文書 3 (“地図”、1) (“閲覧”、1) (“ソフト”、1) (“発売”、1)”が生成され、文書 4 に対応する特徴ベクトル 4 0 4 a として“文書 4 (“地図閲覧ソフト”、1) (“発売”、1)”が生成される。

【 0 0 3 1 】

次に類似文書の検索処理として、種文書特徴ベクトル生成処理ステップ 2 2 0 が実行され、種文書に対応する種文書特徴ベクトル 4 0 7 a が生成される。本図に示した例では、種文書特徴ベクトル 4 0 7 a として、“種文書 (“地図閲覧ソ

フト”、1)” が生成される。

【0032】

そして類似度算出ステップ222において、種文書に対する各登録文書の類似度を算出する。この結果、類似度算出結果408aが出力される。本図に示した例では、数4及び数5に示す様に、文書3の類似度0.000及び文書4の類似度0.710と算出される。

【0033】

【数4】

数4

$$\begin{aligned} & \frac{1 \times 0}{\sqrt{1^2} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2}} \\ = & \frac{0}{2} = 0.000 \end{aligned}$$

【0034】

【数5】

数5

$$\begin{aligned} & \frac{1 \times 1}{\sqrt{1^2} \times \sqrt{1^2 + 1^2}} \\ = & \frac{1}{\sqrt{2}} = 0.710 \end{aligned}$$

【0035】

この結果、文書3の内容は種文書に対して関連があるにも関わらず、従来技術1では文書3の内容は種文書に対して全く類似していないものと算出されてしまう。

【0036】

これは、種文書の特徴ベクトルの要素として抽出される特徴語が複数の単語で構成されているにもかかわらず、最長一致の特徴語「地図閲覧ソフト」のみを特徴ベクトルの要素として類似度算出に用いた為に、特徴語を構成する各単語の持つ個々の概念が類似度に反映されないことによるものである。すなわち、特徴語

を構成する各単語それぞれを含む登録文書に対して類似度が付与されず、検索漏れが発生してしまうことになる。

【0037】

一方、前記の最長一致の特徴語「地図閲覧ソフト」の代わりに、「地図閲覧ソフト」を構成する各単語「地図」「閲覧」「ソフト」を用いることで前記の様な検索漏れを防止することができるが、この場合には「地図閲覧ソフト」とは類似度の低い文書がノイズとして検索される可能性が高くなる。以上が従来技術1の問題点である。

【0038】

本発明の目的は上記問題を解決し、検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を精度良く検索することが可能な技術を提供することにある。本発明の他の目的は検索漏れが無くノイズの少ない高精度な類似文書検索を実現することが可能な技術を提供することにある。

【0039】

【課題を解決するための手段】

本発明は、指定された文書と類似する文書を検索する類似文書検索システムにおいて、複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索するものである。

【0040】

本発明の類似文書検索システムでは、処理対象文書から抽出された特徴語候補が複数の特徴語から構成されている複合特徴語であるか、単一の単語から構成されている単独特徴語であるかを判定し、複合特徴語と判定された場合には複合特徴語及びその複合特徴語を構成する構成特徴語を特徴語として抽出し、単独特徴語と判定された場合には該特徴語そのものを抽出する。

【0041】

すなわち、複合特徴語とその複合特徴語を構成する構成特徴語を抽出し、その抽出した複合特徴語及び構成特徴語を類似度算出に使用することにより、検索漏れの無い高精度な類似文書検索を実現することが可能となる。

【0042】

以上の様に本発明の類似文書検索システムによれば、複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索するので、検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を精度良く検索することが可能である。

【 0 0 4 3 】

【発明の実施の形態】

(実施形態 1)

以下に指定された種文書中の複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索する実施形態 1 の類似文書検索システムについて説明する。

【 0 0 4 4 】

図 1 は本実施形態の類似文書検索システムの概略構成を示す図である。図 1 に示す様に本実施形態の類似文書検索システムは、システム制御処理部 1 1 0 と、登録制御処理部 1 1 1 と、検索制御処理部 1 1 2 と、登録文書取得処理部 1 2 0 と、登録文書特徴ベクトル登録処理部 1 2 1 と、検索条件解析処理部 1 3 0 と、種文書類似度算出処理部 1 3 1 と、検索結果出力処理部 1 3 2 と、登録文書特徴ベクトル読込処理部 1 6 0 と、類似度算出処理部 1 6 1 と、特徴ベクトル生成処理部 1 7 0 と、特徴語抽出処理部 1 7 1 と、文書解析処理部 1 7 2 と、複合特徴語判定処理部 1 7 3 と、出現頻度計数処理部 1 7 4 とを有している。

【 0 0 4 5 】

システム制御処理部 1 1 0 は、キーボード 1 0 1 から入力されたコマンドを解析し、登録制御処理部 1 1 1 または検索制御処理部 1 1 2 を起動する処理部である。登録制御処理部 1 1 1 は、登録文書取得処理部 1 2 0 を起動し、登録対象として指定された文書の特徴ベクトルの磁気ディスク装置 1 0 3 への格納を制御する処理部である。

【 0 0 4 6 】

検索制御処理部 1 1 2 は、検索条件解析処理部 1 3 0、種文書類似度算出処理部 1 3 1、検索結果出力処理部 1 3 2 を起動し、検索条件で指定された種文書に類似する文書の検索を制御する処理部である。

## 【 0 0 4 7 】

登録文書取得処理部 1 2 0 は、登録対象の文書を取得する処理部である。登録文書特徴ベクトル登録処理部 1 2 1 は、登録対象の文書の特徴ベクトルを磁気ディスク装置 1 0 3 へ格納する処理部である。検索条件解析処理部 1 3 0 は、検索条件で指定された種文書を取得する処理部である。

## 【 0 0 4 8 】

種文書類似度算出処理部 1 3 1 は、登録文書特徴ベクトル読込処理部 1 6 0 及び類似度算出処理部 1 6 1 を起動し、種文書から抽出された特徴語と各登録文書との間の類似度を算出する処理部である。検索結果出力処理部 1 3 2 は、前記算出された類似度算出結果を検索結果として出力する処理部である。

## 【 0 0 4 9 】

登録文書特徴ベクトル読込処理部 1 6 0 は、磁気ディスク装置 1 0 3 に格納された登録文書特徴ベクトルファイル 1 8 0 を読み込む処理部である。類似度算出処理部 1 6 1 は、種文書特徴ベクトルに対する登録文書特徴ベクトルのなす角度の余弦を算出し、種文書に対する登録文書の類似度を算出する処理部である。

## 【 0 0 5 0 】

特徴ベクトル生成処理部 1 7 0 は、特徴語抽出処理部 1 7 1 及び出現頻度計数処理部 1 7 4 を起動し、処理対象文書の特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該処理対象文書の特徴語として抽出して処理対象文書の特徴ベクトルを生成する処理部である。

## 【 0 0 5 1 】

特徴語抽出処理部 1 7 1 は、文書解析処理部 1 7 2 及び複合特徴語判定処理部 1 7 3 を起動し、処理対象文書から特徴語または複合特徴語及び構成特徴語を抽出する処理部である。文書解析処理部 1 7 2 は、登録対象の文書である登録文書や所望の検索内容を含んだ種文書等の処理対象文書から特徴語の候補となる特徴語候補を抽出する処理部である。

## 【 0 0 5 2 】

複合特徴語判定処理部 1 7 3 は、前記抽出された特徴語候補に対応する特徴語

にその構成特徴語を示す構成特徴語情報としてそれらの構成特徴語のポインタ情報が登録されている場合に、当該特徴語候補が複合特徴語であると判定する処理部である。出現頻度計数処理部 174 は、処理対象文書から抽出された各特徴語の当該処理対象文書における出現頻度を計数する処理部である。

#### 【0053】

類似文書検索システムをシステム制御処理部 110、登録制御処理部 111、検索制御処理部 112、登録文書取得処理部 120、登録文書特徴ベクトル登録処理部 121、検索条件解析処理部 130、種文書類似度算出処理部 131、検索結果出力処理部 132、登録文書特徴ベクトル読込処理部 160、類似度算出処理部 161、特徴ベクトル生成処理部 170、特徴語抽出処理部 171、文書解析処理部 172、複合特徴語判定処理部 173 及び出現頻度計数処理部 174 として機能させる為のプログラムは、CD-ROM等の記録媒体に記録され磁気ディスク等に格納された後、メモリにロードされて実行されるものとする。なお前記プログラムを記録する記録媒体はCD-ROM以外の他の記録媒体でも良い。また前記プログラムを当該記録媒体から情報処理装置にインストールして使用しても良いし、ネットワークを通じて当該記録媒体にアクセスして前記プログラムを使用するものとしても良い。

#### 【0054】

本実施形態の類似文書検索システムは、ディスプレイ 100、キーボード 101、中央演算処理装置であるCPU 102、磁気ディスク装置 103、フロッピディスクドライブであるFDD 104、主メモリ 105、これらを結ぶバス 106 及び他の機器と本システムを接続するネットワーク 108 から構成される。

#### 【0055】

磁気ディスク装置 103 は二次記憶装置の一つであり、登録文書特徴ベクトルファイル 180 及び特徴語辞書ファイル 181 が格納される。FDD 104 を介してフロッピディスク 107 に格納されている情報が、主メモリ 105 或いは磁気ディスク装置 103 へ読み込まれる。

#### 【0056】

主メモリ 105 には、システム制御処理部 110、登録制御処理部 111、検



索制御処理部 1 1 2、登録文書取得処理部 1 2 0、登録文書特徴ベクトル登録処理部 1 2 1、検索条件解析処理部 1 3 0、種文書類似度算出処理部 1 3 1、検索結果出力処理部 1 3 2 及び共有ライブラリ 1 4 0 が格納されると共にワークエリア 1 4 1 が確保される。共有ライブラリ 1 4 0 には、特徴ベクトル生成処理部 1 7 0、特徴語抽出処理部 1 7 1 及び出現頻度計数処理部 1 7 4 が格納される。

## 【 0 0 5 7 】

種文書類似度算出処理部 1 3 1 は、登録文書特徴ベクトル読込処理部 1 6 0 及び類似度算出処理部 1 6 1 で構成される。特徴ベクトル生成処理部 1 7 0 は、特徴語抽出処理部 1 7 1 及び出現頻度計数処理部 1 7 4 を呼び出す構成をとる。特徴語抽出処理部 1 7 1 は、文書解析処理部 1 7 2 及び複合特徴語判定処理部 1 7 3 で構成される。

## 【 0 0 5 8 】

登録制御処理部 1 1 1 及び検索制御処理部 1 1 2 は、キーボード 1 0 1 からのユーザによる指示に応じてシステム制御処理部 1 1 0 によって起動され、それぞれ登録文書取得処理部 1 2 0、特徴ベクトル生成処理部 1 7 0 及び登録文書特徴ベクトル登録処理部 1 2 1 の制御と、検索条件解析処理部 1 3 0、特徴ベクトル生成処理部 1 7 0、種文書類似度算出処理部 1 3 1 及び検索結果出力処理部 1 3 2 の制御を行なう。

## 【 0 0 5 9 】

なお本実施形態では、キーボード 1 0 1 から入力されたコマンドにより、登録制御処理部 1 1 1 や検索制御処理部 1 1 2 が起動されるものとしたが、他の入力装置を介して入力されたコマンド或いはイベントにより起動されるものであっても構わない。

## 【 0 0 6 0 】

また、本実施形態の類似文書検索システムをこれらの処理部として機能させる為のプログラムは、磁気ディスク装置 1 0 3、フロッピディスク 1 0 7、或いは MO、CD-ROM、DVD 等の記録媒体（図 1 には示していない）に格納され、駆動装置を介して主メモリ 1 0 5 に読み込まれ、CPU 1 0 2 によって実行されるものとするが、これらのプログラムをネットワーク 1 0 8 を介して主メモリ

105に読み込み、CPU102によって実行することも同様に可能である。

【0061】

更に、本実施形態では登録文書特徴ベクトルファイル180及び特徴語辞書ファイル181を磁気ディスク装置103に格納するものとしたが、フロッピディスク107、MO、CD-ROM、DVD等の記録媒体（図1には示していない）に格納し、駆動装置を介して主メモリ105に読み込み利用することも可能である。また、これらのファイルはネットワーク108を介して、他のシステムに接続された記録媒体（図1には示していない）に格納されるものとしても良いし、或いはネットワーク108に直接接続された記録媒体に格納されるものとしても構わない。

【0062】

以下、本実施形態における類似文書検索システムの処理手順について説明する。

図本実施形態のシステム制御処理部110の処理内容を示す図である。まず、システム制御処理部110の処理手順について図2のPAD図を用いて説明する。

【0063】

システム制御処理部110は、まずステップ800で、キーボード101から入力されたコマンドを解析する。そしてステップ801で、この結果が登録実行のコマンドであると解析された場合には、ステップ802で登録制御処理部111を起動して文書の登録を行なう。またステップ801で、検索実行のコマンドであると解析された場合には、ステップ803で検索制御処理部112を起動して、類似文書の検索を行なう。以上が、システム制御処理部110の処理手順である。

【0064】

図3は本実施形態の登録制御処理部111の処理内容を示す図である。図2に示したシステム制御処理部110のステップ802で起動される登録制御処理部111の処理手順について、図3のPAD図を用いて説明する。

【0065】

登録制御処理部 1 1 1 では、まずステップ 9 0 0 において登録文書取得処理部 1 2 0 を起動し、登録対象として指定された文書（以下、登録対象文書と呼ぶ）を読み込み、ワークエリア 1 4 1 に格納する。

【 0 0 6 6 】

次に、ステップ 9 0 1 において、共有ライブラリ 1 4 0 に格納されている特徴ベクトル生成処理部 1 7 0 を起動し、登録対象文書に対する特徴ベクトルを生成し、ワークエリア 1 4 1 に格納する。

【 0 0 6 7 】

そして、ステップ 9 0 2 において、登録文書特徴ベクトル登録処理部 1 2 1 を起動し、ワークエリア 1 4 1 に格納されている登録文書特徴ベクトルを磁気ディスク装置 1 0 3 へ格納する。以上が、登録制御処理部 1 1 1 の処理手順である。

【 0 0 6 8 】

図 4 は本実施形態の特徴ベクトル生成処理部 1 7 0 の処理内容を示す図である。図 3 に示した登録制御処理部 1 1 1 のステップ 9 0 1 で起動される特徴ベクトル生成処理部 1 7 0 の処理手順について、図 4 の P A D 図を用いて説明する。

【 0 0 6 9 】

特徴ベクトル生成処理部 1 7 0 では、まずステップ 1 0 0 0 において特徴語抽出処理部 1 7 1 を起動し、ワークエリア 1 4 1 に格納された処理対象文書から特徴語を抽出する。次に、ステップ 1 0 0 1 において、出現頻度計数処理部 1 7 4 を起動し、ワークエリア 1 4 1 に格納された各特徴語の処理対象文書における出現頻度を計数する。以上が、特徴ベクトル生成処理部 1 7 0 の処理手順である。なお、本特徴ベクトル生成処理部 1 7 0 は共有ライブラリ 1 4 0 に格納されており、後述する文書検索処理における検索制御処理部 1 1 2 から実行され、種文書に対する特徴ベクトルの生成においても使用される。

【 0 0 7 0 】

図 5 は本実施形態の特徴語抽出処理部 1 7 1 の処理内容を示す図である。図 4 に示した特徴ベクトル生成処理部 1 7 0 のステップ 1 0 0 0 で起動される特徴語抽出処理部 1 7 1 の処理手順について、図 5 の P A D 図を用いて説明する。

【 0 0 7 1 】

特徴語抽出処理部 171 は、まずステップ 1400 において文書解析処理部 172 を起動し、ワークエリア 141 に格納された処理対象文書中の文字列と特徴語辞書ファイル 181 中の特徴語とを比較し、特徴語辞書ファイル 181 中の特徴語と一致する文字列を特徴語候補として処理対象文書から抽出する。

【0072】

次にステップ 1401 において、複合特徴語判定処理部 173 を起動し、上記ステップ 1400 において抽出された特徴語候補に対応する特徴語辞書ファイル 181 中の特徴語に構成特徴語の格納位置を示すポインタ情報が登録されているかどうかを調べ、特徴語辞書ファイル 181 中の特徴語に前記ポインタ情報が登録されている場合には、当該特徴語候補が複合特徴語であると判定する。

【0073】

そして、特徴語候補が複合特徴語であると判定された場合には、ステップ 1402 を実行し、前記ポインタ情報で示された特徴語をその複合特徴語の構成特徴語として読み出して、それらの複合特徴語及び構成特徴語を処理対象文書の特徴語として抽出する。

【0074】

前記の様に本実施形態では、抽出された特徴語候補に対応する特徴語辞書ファイル 181 中の特徴語に構成特徴語の格納位置を示すポインタ情報が登録されているかどうかを調べることにより、当該特徴語候補が複合特徴語であるかを判定し、前記ポインタ情報を用いて構成特徴語の読み出しを行なうので、特徴語候補が複合特徴語であるかの判定及び構成特徴語の読み出しを高速に行なうことが可能である。なお、本実施形態では複合特徴語に構成特徴語を示すポインタ情報を格納するものとしたが、複合特徴語内に分割位置を格納しておくものとしても良いし、構成特徴語そのものを格納しておくものとしても良い。

【0075】

また、ステップ 1401 において、特徴語候補が複合特徴語でないと判定された場合にはステップ 1403 を実行し、特徴語候補そのものを処理対象文書の特徴語として抽出する。以上が、特徴語抽出処理部 171 の処理手順である。

【0076】

図 6 は本実施形態の検索制御処理部 1 1 2 の処理内容を示す図である。図 2 に示したシステム制御処理部 1 1 0 のステップ 8 0 3 で起動される検索制御処理部 1 1 2 の処理手順について、図 6 の P A D 図を用いて説明する。

【 0 0 7 7 】

検索制御処理部 1 1 2 は、まずステップ 1 1 0 0 において、検索条件解析処理部 1 3 0 を起動し、検索条件で指定された種文書を取得する。そしてステップ 1 1 0 1 において、共有ライブラリ 1 4 0 に格納された特徴ベクトル生成処理部 1 7 0 を起動し、上記ステップ 1 1 0 0 で取得された種文書に対する種文書特徴ベクトルを生成する。

【 0 0 7 8 】

次にステップ 1 1 0 2 において、種文書類似度算出処理部 1 3 1 を起動し、種文書に対する各登録文書の類似度を算出する。そしてステップ 1 1 0 3 において、検索結果出力処理部 1 3 2 を起動し、上記ステップ 1 1 0 1 で算出された類似度算出結果を検索結果として出力する。

【 0 0 7 9 】

ここで、検索結果の出力先は、ディスプレイ 1 0 0 に表示するものとしても良いし、ワークエリア 1 4 1 や磁気ディスク装置 1 0 3 上に格納するものとしても良い。また、類似度算出結果をディスプレイ 1 0 0 に出力する場合には、類似度の降順に出力するものとしても良いし、文書に付与された管理番号の昇順或いは降順に出力するものとしても良い。以上が検索制御処理部 1 1 2 の処理手順である。

【 0 0 8 0 】

図 7 は本実施形態の種文書類似度算出処理部 1 3 1 の処理内容を示す図である。図 6 に示した検索制御処理部 1 1 2 のステップ 1 1 0 2 で起動される種文書類似度算出処理部 1 3 1 の処理手順について、図 7 の P A D 図を用いて説明する。

【 0 0 8 1 】

種文書類似度算出処理部 1 3 1 は、まずステップ 1 3 0 0 において、登録文書特徴ベクトル読込処理部 1 6 0 を起動し、磁気ディスク装置 1 0 3 に格納された登録文書特徴ベクトルファイル 1 8 0 を読み込み、ワークエリア 1 4 1 に格納す

る。

#### 【0082】

そしてステップ1301において、ワークエリア141に格納された全ての登録文書特徴ベクトルに対して、ステップ1302を繰り返し実行する。ステップ1302では、類似度算出処理部161を起動し、種文書特徴ベクトルに対する登録文書特徴ベクトルのなす角度の余弦を算出し、種文書に対する登録文書の類似度としてワークエリア141に格納する。以上が種文書類似度算出処理部131の処理手順である。

#### 【0083】

以下、本実施形態における類似文書検索システムの具体的な処理手順を図8～図11を用いて説明する。まず、本実施形態における類似文書検索システムにおける文書の登録処理について、図8を用いて説明する。

#### 【0084】

図8は本実施形態の文書の登録処理の処理内容を示す図である。図8では、文書3「地図情報閲覧ソフトを開発、発売したA社は、・・・」及び文書4「多くの地図閲覧ソフトが発売されているが、・・・」が文書データベースに登録される場合の処理の流れを表している。

#### 【0085】

まず、本実施形態の類似文書検索システムにおいて、登録文書取得処理部120は、登録対象の文書3及び文書4を読み込み、ワークエリア141に格納する。次に特徴ベクトル生成処理部170は、登録対象の文書3及び文書4に対応する登録文書特徴ベクトル403a及び404bを作成し、ワークエリア141に格納する。そして、登録文書特徴ベクトル登録処理部121は、ワークエリア141上の登録文書特徴ベクトルを登録文書特徴ベクトルファイル180に格納する。以上が、本実施形態に示した類似文書検索システムにおける文書の登録処理である。

#### 【0086】

次に、本実施形態における類似文書検索システムにおける類似文書の検索処理について、図9を用いて説明する。

## 【0087】

図9は本実施形態の類似文書の検索処理の処理内容を示す図である。図9では、種文書「最新の地図閲覧ソフトについて」が入力された場合の例を表している。まず、検索条件解析処理部130は、検索条件で指定された種文書を取得し、ワークエリア141に格納する。

## 【0088】

そして、特徴ベクトル生成処理部170は、ワークエリア141に格納された種文書に対応する種文書特徴ベクトル407bを生成し、ワークエリア141に格納する。

## 【0089】

次に、登録文書特徴ベクトル読込処理部160は、前記文書の登録処理で作成された登録文書特徴ベクトルファイル180を読み込み、登録文書特徴ベクトル403a及び404bをワークエリア141に格納する。

## 【0090】

【数6】

数6

$$\frac{1 \times 0 + 1 \times 1 + 1 \times 1 + 1 \times 1}{\sqrt{1^2 + 1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2}} = \frac{3}{2 \times 2} = 0.750$$

## 【0091】

【数7】

数7

$$\frac{1 \times 1 + 1 \times 1 + 1 \times 1 + 1 \times 1}{\sqrt{1^2 + 1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2}} = \frac{4}{2 \sqrt{5}} = 0.894$$

## 【0092】

そして、類似度算出処理部161は、前記ステップ170で生成された種文書

特徴ベクトル407bと登録文書特徴ベクトル403a及び404bのなす角度の余弦を数6及び数7の様に算出し、種文書に対する登録文書の類似度算出結果408bを出力する。以上が、本実施形態における類似文書検索システムにおける類似文書の検索処理手順である。

#### 【0093】

次に、本実施形態における類似文書検索システムにおける特徴ベクトルの生成処理手順について図10を用いて説明する。

#### 【0094】

図10は本実施形態の特徴ベクトルの生成処理の処理内容を示す図である。図10では、種文書「最新の地図閲覧ソフトについて」が入力された場合にその特徴ベクトルが作成される手順を表している。

#### 【0095】

まず、文書解析処理部172は、ワークエリア141に格納された処理対象文書である種文書1601“最新の地図閲覧ソフトについて”中の文字列と特徴語辞書ファイル181中の特徴語とを比較し、特徴語辞書ファイル181中の特徴語と一致する文字列“地図閲覧ソフト”を特徴語候補1602として種文書1601から抽出する。

#### 【0096】

そして、複合特徴語判定処理部173は、特徴語辞書ファイル181中の特徴語“地図閲覧ソフト”に構成特徴語の格納位置を示すポインタ情報が登録されているかどうかを調べ、特徴語候補1602“地図閲覧ソフト”が複数の特徴語で構成される複合特徴語であるかを判定する。この結果、特徴語候補1602“地図閲覧ソフト”は複数の特徴語“地図”、“閲覧”、“ソフト”から構成されるものと判定され、複合特徴語と判定される。

#### 【0097】

次に、特徴語抽出処理部171は、上記複合特徴語判定処理部173の結果、複合特徴語と判定された“地図閲覧ソフト”から、これを構成する特徴語1604“地図”“閲覧”“ソフト”を前記ポインタ情報により抽出する。そして、出現頻度計数処理部174は、上記特徴語抽出処理部171で抽出された各特徴語



について、種文書 1 6 0 1 内での出現頻度を計数し、特徴語とその出現頻度を特徴ベクトル 1 6 0 5 として出力する。以上が、本実施形態における類似文書検索システムにおける特徴ベクトルの生成処理手順である。

## 【 0 0 9 8 】

以上説明した様に本実施形態によれば、複合特徴語だけでなく、複合特徴語を構成する構成特徴語を特徴ベクトルの要素として類似度算出に使用する。この結果として、最長一致の様に“地図閲覧ソフト”を含むノイズの少ない類似文書検索を行なうと共に検索漏れの無い高精度な類似文書検索を実現することができる。

## 【 0 0 9 9 】

なお本実施形態では、登録対象文書や種文書を文書としたが、文章或いは文字列であっても構わない。また、本実施形態における特徴ベクトル生成処理では、処理対象中に出現する複合特徴語から複合特徴語及び複合特徴語に含まれる構成特徴語を全て抽出するものとして説明したが、全ての構成特徴語を抽出するのではなく一部を抽出するものとしても構わない。この場合、抽出する構成特徴語の指定方法としては、従来技術 1 の参照文献等に記載されている IDF (Inverted Document Frequency) が予め定められた閾値を越えるものだけを抽出するものとしても良いし、複合特徴語の中で先頭或いは末尾等の予め定められた位置を構成する特徴語だけを抽出するものとしても良い。

## 【 0 1 0 0 】

また本実施形態では、登録対象文書に対する特徴ベクトルを予め作成しておくものとしたが、文書の登録時には全文検索用インデックスを作成しておき、検索時に該当する全文検索用インデックスを参照することにより各登録対象文書に出現頻度を求め、類似度の算出を行なうものとしても良い。更に本実施形態では、特徴語の抽出に特徴語辞書を参照するものとして説明したが、辞書を用いずに特徴語を抽出する技術等を用いることも可能である。

## 【 0 1 0 1 】

また、本実施形態では日本語における類似文書検索システムの例を説明したが、日本語に限らず他言語であっても構わない。すなわち、前述の日本語における

類似文書検索システムの場合には、種文書中に存在する複合特徴語及び該複合特徴語を構成する単語を類似度算出に使用することで検索漏れの無い類似文書検索を実現していたが、例えば英語等の様に単語の境界が明確な言語の場合には、複数の単語の組（一般にフレーズや熟語と呼ばれる）を複合特徴語として取り扱い、フレーズや熟語を用いた検索を行なう際に、本実施形態を適用することが可能となる。

【0102】

これにより、他言語においても意味のつながりのある単語の組の内容を考慮した類似度算出を行なうことができる様になり、検索漏れの少ない多言語対応の類似文書検索を提供することができる様になる。

【0103】

まず、従来技術1を英文対応類似文書検索システムに適用した場合の問題点について図11を用いて説明する。

【0104】

図11は従来技術1を英文対応類似文書検索システムに適用した場合の問題点を示す図である。本図では、文書5「This juice is made of carrot...」及び文書6「— Carrot Juice — 1. Cut carrot into some pieces...」が登録された文書データベースに対して、種文書「How to make carrot juice」が入力された場合の例を表している。

【0105】

まず文書の登録処理として、ステップ210が実行され、各文書に対応する登録文書特徴ベクトル1702及び1703が生成される。本図に示した例では、文書5に対応する登録文書特徴ベクトル1702として“文書5 (“carrot”,1) (“juice”,1)”が生成され、文書6に対応する登録文書特徴ベクトル1703として“文書6 (“carrot juice”,1) (“carrot”,1)”が生成される。

【0106】

次に類似文書の検索処理として、ステップ220が実行され、種文書に対応する種文書特徴ベクトル1706が生成される。本図に示した例では、種文書特徴ベクトル1706として、“種文書 (“carrot juice”,1)”が生成される。

【0107】

そしてステップ222において、種文書に対する各登録文書の類似度を算出する。この結果、類似度算出結果1707が出力される。本図に示した例では、数8及び数9に示す様に、文書5の類似度0.000及び文書6の類似度0.710と算出される。

【0108】

【数8】

数8

$$\frac{1 \times 0}{\sqrt{1^2} \times \sqrt{1^2 + 1^2}} = \frac{0}{\sqrt{2}} = 0.000$$

【0109】

【数9】

数9

$$\frac{1 \times 1}{\sqrt{1^2} \times \sqrt{1^2 + 1^2}} = \frac{1}{\sqrt{2}} = 0.710$$

【0110】

この結果、文書5の内容は種文書に対して関連があるにも関わらず、従来技術1では文書5の内容は種文書に対して全く類似していないものと算出されてしまう。

【0111】

これは、種文書の特徴ベクトルの要素として抽出される特徴語が複数の単語の組で構成されているにもかかわらず、該特徴語のみを特徴ベクトルの要素として類似度算出に用いた為に、特徴語を構成する各単語の持つ個々の概念が類似度に反映されないことによるものである。

【0112】

すなわち、ノイズ等を減らす為に複数の単語の組である"carrot juice"等の特徴語とした場合には、"carrot juice"を含む文書6に対する検索精度が向上し、"carrot juice"を含まない登録文書は検索されなくなるが、その特徴語を構成する各単語それぞれを含む登録文書の文書5に対して類似度が付与されず、検索漏れが発生してしまうことになる。

【0113】

以上説明した様に従来技術1を英文対応類似文書検索システムに適用した場合にも、日本語の場合と同様の問題が生じてしまうことになる。上記問題に対し、本実施形態を英文対応類似文書検索システムに適用することにより、日本語の場合と同様に解決することができるようになる。

【0114】

以下、図12に本実施形態を適用した英文対応類似文書検索システムの処理概要を示す。

【0115】

図12は本実施形態の英文対応類似文書検索システムの処理概要を示す図である。図12は、文書5「This juice is made of carrot...」及び文書6「— Carrot juice — 1. Cut carrot into some pieces...」が登録された文書データベースに対して、種文書「How to make carrot juice」が入力された場合の例を表している。

【0116】

まず文書の登録処理のステップ210では、各文書に対応する登録文書特徴ベクトル1702及び1703を生成する。本図に示した例では、文書5に対応する登録文書特徴ベクトル1702として“文書5 ("carrot",1) ("juice",1)”を生成し、文書6に対応する登録文書特徴ベクトル1703aとして“文書6 ("carrot juice",1) ("carrot",2) ("juice",1)”を生成する。

【0117】

次に文書の検索処理のステップ220では、種文書に対応する種文書特徴ベクトル1706aを生成する。本図に示した例では、種文書特徴ベクトル1706aとして、“種文書 ("carrot juice",1) ("carrot",1) ("juice",1)”を生

成する。

【0 1 1 8】

そしてステップ2 2 2において、種文書に対する各登録文書の類似度を算出する。この結果、類似度算出結果1 7 0 7 aを出力する。数1 0及び数1 1に示す様に、本図に示した例では文書5の類似度0. 8 1 6及び文書6の類似度0. 9 4 3と算出される。

【0 1 1 9】

【数1 0】

数10

$$\frac{1 \times 0 + 1 \times 1 + 1 \times 1}{\sqrt{1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 1^2}} = \frac{2}{\sqrt{3} \times \sqrt{2}} = 0.816$$

【0 1 2 0】

【数1 1】

数11

$$\frac{1 \times 1 + 1 \times 2 + 1 \times 1}{\sqrt{1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 2^2 + 1^2}} = \frac{4}{\sqrt{3} \times \sqrt{6}} = 0.943$$

【0 1 2 1】

以上が本実施形態を適用した英文対応類似文書検索システムの処理概要である。前記の様に本実施形態を適用した英文対応類似文書検索システムにおいても複合特徴語を考慮することにより、従来技術1では検索することができない文書5を検索することができるようになる。

【0 1 2 2】

以上説明した様に本実施形態の類似文書検索システムによれば、複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索するので、検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を

精度良く検索することが可能である。

【 0 1 2 3 】

(実施形態 2)

以下に複合特徴語から抽出された構成特徴語の登録文書内での出現距離を考慮した重み付けを行なう実施形態 2 の類似文書検索システムについて説明する。

【 0 1 2 4 】

本実施形態を適用した類似文書検索システムでは、複合特徴語から抽出された構成特徴語の登録文書内での出現距離を考慮した重み付けを行なうものであり、種文書の同一複合特徴語から抽出された関連性の高い構成特徴語が、関連の高い出現関係にある登録文書に対して高い類似度を付与することにより、より内容の近い登録文書を検索し、高精度な検索結果を得ることができるようになる。

【 0 1 2 5 】

図 1 3 は本実施形態の特徴ベクトル生成処理部 1 7 0 a の構成を示す図である。図 1 3 に示す様に本実施形態の類似文書検索システムは出現位置取得処理部 1 9 0 0 を有している。出現位置取得処理部 1 9 0 0 は、特徴語抽出処理部 1 7 1 で抽出された各特徴語について、処理対象文書内での出現位置を取得する処理部である。

【 0 1 2 6 】

類似文書検索システムを出現位置取得処理部 1 9 0 0 として機能させる為のプログラムは、CD-ROM等の記録媒体に記録され磁気ディスク等に格納された後、メモリにロードされて実行されるものとする。なお前記プログラムを記録する記録媒体はCD-ROM以外の他の記録媒体でも良い。また前記プログラムを当該記録媒体から情報処理装置にインストールして使用しても良いし、ネットワークを通じて当該記録媒体にアクセスして前記プログラムを使用するものとしても良い。

【 0 1 2 7 】

図 1 4 は本実施形態の種文書類似度算出処理部 1 3 1 a の構成を示す図である。図 1 4 に示す様に本実施形態の類似文書検索システムは重み係数算出処理部 2 0 0 0 を有している。重み係数算出処理部 2 0 0 0 は、種文書から抽出された構

成特徴語に一致する登録文書の特徴語について、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を算出する処理部である。

#### 【 0 1 2 8 】

類似文書検索システムを重み係数算出処理部 2 0 0 0 として機能させる為のプログラムは、CD-ROM等の記録媒体に記録され磁気ディスク等に格納された後、メモリにロードされて実行されるものとする。なお前記プログラムを記録する記録媒体はCD-ROM以外の他の記録媒体でも良い。また前記プログラムを当該記録媒体から情報処理装置にインストールして使用しても良いし、ネットワークを通じて当該記録媒体にアクセスして前記プログラムを使用するものとしても良い。

#### 【 0 1 2 9 】

本実施形態は、実施形態 1（図 1）とほぼ同様の構成を取るが、特徴ベクトル生成処理部 1 7 0 及び種文書類似度算出処理部 1 3 1 の構成が異なる。特徴ベクトル生成処理部 1 7 0 a では、図 1 3 に示す様に、出現位置取得処理部 1 9 0 0 が用いられる。また、種文書類似度算出処理部 1 3 1 a では、図 1 4 に示す様に重み係数算出処理部 2 0 0 0 が用いられる。

#### 【 0 1 3 0 】

以下、本実施形態における処理手順の内、まず実施形態 1 とは異なる特徴ベクトル生成処理部 1 7 0 a の処理手順について、図 1 5 に示す PAD 図を用いて説明する。

#### 【 0 1 3 1 】

図 1 5 は本実施形態の特徴ベクトル生成処理部 1 7 0 a の処理内容を示す図である。ここで、実施形態 1 における特徴ベクトル生成処理部 1 7 0（図 4）と異なる点は、出現位置取得ステップ 2 1 0 0 が加わるだけである。他の処理ステップの処理手順は、実施形態 1 で説明した通りである。

#### 【 0 1 3 2 】

出現位置取得ステップ 2 1 0 0 では、出現位置取得処理部 1 9 0 0 を起動し、ワークエリア 1 4 1 に格納された各単語の、処理対象文書における出現位置を取得する。以上が、特徴ベクトル生成処理部 1 7 0 a の処理手順である。

## 【0133】

次に、本実施形態における種文書類似度算出処理部131aの処理手順について、図16に示すPAD図を用いて説明する。

## 【0134】

図16は本実施形態の種文書類似度算出処理部131aの処理内容を示す図である。ここで、実施形態1における種文書類似度算出処理部131（図7）と異なる点は、重み係数算出ステップ2200が加わるだけである。他の処理ステップの処理手順は、実施形態1で説明した通りである。

## 【0135】

重み係数算出ステップ2200では、重み係数算出処理部2000を起動し、種文書特徴ベクトルの各要素の内、同一の複合特徴語から抽出された構成特徴語の組に対して重み係数を算出し、種文書特徴ベクトルの要素に乗じる。以上が、種文書類似度算出処理部131aの処理手順である。

## 【0136】

以下、本実施形態における類似文書検索システムの具体的な処理手順を図17～図19を用いて説明する。まず、本実施形態における類似文書検索システムにおける文書の登録処理について、図17を用いて説明する。

## 【0137】

図17は本実施形態の文書登録処理の概要を示す図である。図17では、文書3「地図情報閲覧ソフトを開発、発売したA社は、・・・」及び文書4「多くの地図閲覧ソフトが発売されているが、・・・」が文書データベースに登録される場合の処理の流れを表している。

## 【0138】

まず登録文書取得処理部120は、文書3及び文書4を読み込み、ワークエリア141に格納する。次に特徴ベクトル生成処理部170aは、登録対象の文書3及び文書4に対して対応する登録文書特徴ベクトル2300及び2301を作成し、ワークエリア141に格納する。

## 【0139】

本図に示した例では、文書3に対応する登録文書特徴ベクトル2300として



“文書 3 (“地図”, 1) [1]、 (“閲覧”, 1) [5]、 (“ソフト”, 1) [7]、 (“発売”, 1) [14]” が生成され、 文書 4 に対応する登録文書特徴ベクトル 2301 として “文書 4 (“地図閲覧ソフト”, 1) [4]、 (“地図”, 1) [4]、 (“閲覧”, 1) [6]、 (“ソフト”, 1) [8]、 (“発売”, 1) [12]” が生成される。 なお、 ここで “ (“地図”, 1) [1]” の丸括弧 ( ) 内は特徴語 “地図” が 1 回出現することを表し、 角括弧 [ ] 内の “1” は特徴語 “地図” の文字位置が 1 であることを表している。

【0140】

そして、 登録文書特徴ベクトル登録処理部 121 は、 ワークエリア 141 上の登録文書特徴ベクトルを登録文書特徴ベクトルファイル 180 として格納する。 以上が、 本実施形態に示した類似文書検索システムにおける文書の登録処理である。

【0141】

次に、 本実施形態における類似文書検索システムにおける類似文書の検索処理について、 図 18 を用いて説明する。

【0142】

図 18 は本実施形態の類似文書の検索処理の処理内容を示す図である。 図 18 では、 種文書「最新の地図閲覧ソフトについて」が入力された場合の例を表している。 まず、 検索条件解析処理部 130 は、 検索条件で指定された種文書を取得し、 ワークエリア 141 に格納する。

【0143】

そして、 特徴ベクトル生成処理部 170a は、 ワークエリア 141 に格納された種文書に対応する種文書特徴ベクトル 2400 を生成し、 ワークエリア 141 に格納する。

【0144】

次に、 登録文書特徴ベクトル読込処理部 160 は、 前記文書の登録処理で作成された登録文書特徴ベクトルファイル 180 を読み込み、 登録文書特徴ベクトル 2300 及び 2301 をワークエリア 141 に格納する。

【0145】

そして、 重み係数算出処理部 2000 は、 種文書特徴ベクトル 2400 の各要

素が構成特徴語であるかを判定し、該要素がある複合特徴語の構成特徴語である場合には数12に基づいて重みを算出し、重み係数2401として出力する。

【0146】

【数12】

数12

種文書特徴ベクトルの構成特徴語Aに対する重み係数

$$= 1 - \frac{\min(\text{同親構成特徴語との最短距離}, C)}{C}$$

【0147】

本図では、数12（定数 $C=20$ とした）に基づき各要素の重みを算出した場合の例を表している。すなわち、種文書特徴ベクトル2400の要素“地図”は複合特徴語“地図閲覧ソフト”の構成特徴語であるから、同じ複合特徴語“地図閲覧ソフト”から抽出された他の構成特徴語（以下、同親構成特徴語と呼ぶ）である“閲覧”、“ソフト”との最短距離を取得する。本図に示した例では文書3において、“地図”に対する“閲覧”、“ソフト”の最短距離は“4”であるから、重み係数“0.80”が算出されている。

【0148】

そして、類似度算出処理部161において、前記ステップ170aで生成された種文書特徴ベクトル2400と登録文書特徴ベクトル2300及び2301のなす角度の余弦が数13及び数14の様に算出され、種文書に対する登録文書の類似度算出結果2402が出力される。

【0149】

【数13】

数13

$$\frac{1 \times 1 \times 0 + 0.8 \times 1 \times 1 + 0.9 \times 1 \times 1 + 0.9 \times 1 \times 1}{\sqrt{(1 \times 1)^2 + (0.8 \times 1)^2 + (0.9 \times 1)^2 + (0.9 \times 1)^2}} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2}$$

$$= \frac{2.6}{\sqrt{3.26} \sqrt{4}} = 0.720$$

【0150】

【数 1 4】

数14

$$\begin{aligned}
 & \frac{1 \times 1 \times 1 + 0.9 \times 1 \times 1 + 0.9 \times 1 \times 1 + 0.9 \times 1 \times 1}{\sqrt{(1 \times 1)^2 + (0.9 \times 1)^2 + (0.9 \times 1)^2 + (0.9 \times 1)^2}} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2} \\
 &= \frac{3.7}{\sqrt{3.43} \sqrt{5}} = 0.893
 \end{aligned}$$

【0 1 5 1】

以上が、本実施形態における類似文書検索システムにおける類似文書の検索処理手順である。

【0 1 5 2】

次に、本実施形態における類似文書検索システムにおける特徴ベクトルの生成処理手順について図 1 9 を用いて説明する。

【0 1 5 3】

図 1 9 は本実施形態の特徴ベクトルの生成処理の処理内容を示す図である。図 1 9 では、種文書「最新の地図閲覧ソフトについて」が入力された場合の例に、検索特徴ベクトルが作成される手順を表している。

【0 1 5 4】

まず、文書解析処理部 1 7 2 は、ワークエリア 1 4 1 に格納された処理対象文書である種文書 1 6 0 1 “最新の地図閲覧ソフトについて” から特徴語候補 1 6 0 2 “地図閲覧ソフト” を抽出する。

【0 1 5 5】

そして、複合特徴語判定処理部 1 7 3 は、特徴語候補 1 6 0 2 “地図閲覧ソフト” が複数の特徴語で構成される特徴語かを判定する。この結果、特徴語候補 1 6 0 2 “地図閲覧ソフト” は複数の特徴語“地図”、“閲覧”、“ソフト”から構成されるものと判定され、複合特徴語と判定される。

【0 1 5 6】

次に、特徴語抽出処理部 1 7 1 では、上記複合特徴語判定処理部 1 7 3 の結果、複合特徴語と判定された“地図閲覧ソフト”から、これを構成する特徴語 1 6 0 4 “地図”、“閲覧”、“ソフト”を抽出する。そして、出現頻度計数処理部

174 は、上記特徴語抽出処理で抽出された各特徴語について、種文書1601内での出現頻度を計数する。

【0157】

そして、出現位置取得処理部1900は、上記特徴語抽出処理部171で抽出された各特徴語について、種文書1601内での出現位置を取得し、特徴ベクトル2500として出力する。以上が、本実施形態における類似文書検索システムにおける特徴ベクトルの生成処理手順である。

【0158】

以上説明した様に、本実施形態によれば、種文書から抽出された複合特徴語の構成特徴語間の距離を考慮することにより、登録文書内での単語間の関係を考慮した高精度な類似度算出を行なうことができる。すなわち、複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索することにより、検索漏れの無い高精度な類似文書検索が可能となるが、その際に構成特徴語間の距離を考慮して重み付けを行なうことにより、種文書との関連が低い登録文書の類似度を下げて検索時のノイズを削減することが可能である。

【0159】

なお、本実施形態における特徴ベクトル生成処理部170aでは、複合特徴語及び複合特徴語から抽出された構成特徴語の両方を特徴語として抽出していたが、構成特徴語だけを抽出するものとして良い。この場合、重み係数算出や類似度算出に使用される特徴語の要素数が削減される為、より高速な検索を実現することができる。

【0160】

また、本実施形態における特徴ベクトル生成処理部170aでは、各特徴語の出現位置取得処理部1900を出現頻度計数処理部174の後に実施するものとしたが、種文書解析処理部172の実施時に各特徴語候補を抽出するのに合わせて、各特徴語候補文字列の出現位置を抽出しておくものとしても良い。

【0161】

更に、本実施形態における特徴ベクトル2500では、各要素に対応して出現頻度及び出現位置を格納するものとしたが、種文書に対する特徴ベクトル作成処

理では同親構成特徴語をまとめて一つの要素として管理するものとしても良い。  
この様にすることにより、重み係数算出処理時に各要素が構成特徴語か否かを判断する必要がない為、より高速な検索を実現することができる。

【0162】

以上説明した様に本実施形態の類似文書検索システムによれば、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を乗じた類似度を算出するので、検索漏れが無くノイズの少ない高精度な類似文書検索を実現することが可能である。

【0163】

【発明の効果】

本発明によれば複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索するので、検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を精度良く検索することが可能である。

【図面の簡単な説明】

【図1】

実施形態1の類似文書検索システムの概略構成を示す図である。

【図2】

実施形態1のシステム制御処理部110の処理内容を示す図である。

【図3】

実施形態1の登録制御処理部111の処理内容を示す図である。

【図4】

実施形態1の特徴ベクトル生成処理部170の処理内容を示す図である。

【図5】

実施形態1の特徴語抽出処理部171の処理内容を示す図である。

【図6】

実施形態1の検索制御処理部112の処理内容を示す図である。

【図7】

実施形態1の種文書類似度算出処理部131の処理内容を示す図である。

【図8】

実施形態 1 の文書の登録処理の処理内容を示す図である。

【図 9】

実施形態 1 の類似文書の検索処理の処理内容を示す図である。

【図 10】

実施形態 1 の特徴ベクトルの生成処理の処理内容を示す図である。

【図 11】

従来技術 1 を英文対応類似文書検索システムに適用した場合の問題点を示す図である。

【図 12】

実施形態 1 の英文対応類似文書検索システムの処理概要を示す図である。

【図 13】

実施形態 2 の特徴ベクトル生成処理部 170 a の構成を示す図である。

【図 14】

実施形態 2 の種文書類似度算出処理部 131 a の構成を示す図である。

【図 15】

実施形態 2 の特徴ベクトル生成処理部 170 a の処理内容を示す図である。

【図 16】

実施形態 2 の種文書類似度算出処理部 131 a の処理内容を示す図である。

【図 17】

実施形態 2 の文書登録処理の概要を示す図である。

【図 18】

実施形態 2 の類似文書の検索処理の処理内容を示す図である。

【図 19】

実施形態 2 の特徴ベクトルの生成処理の処理内容を示す図である。

【図 20】

従来技術 1 の処理手順の一例を示す図である。

【図 21】

従来技術 1 における特徴ベクトル生成処理の一例を示す図である。

【図 22】

従来技術 1 の概要を示す図である。

【図 2 3】

従来技術 1 の問題点を示す図である。

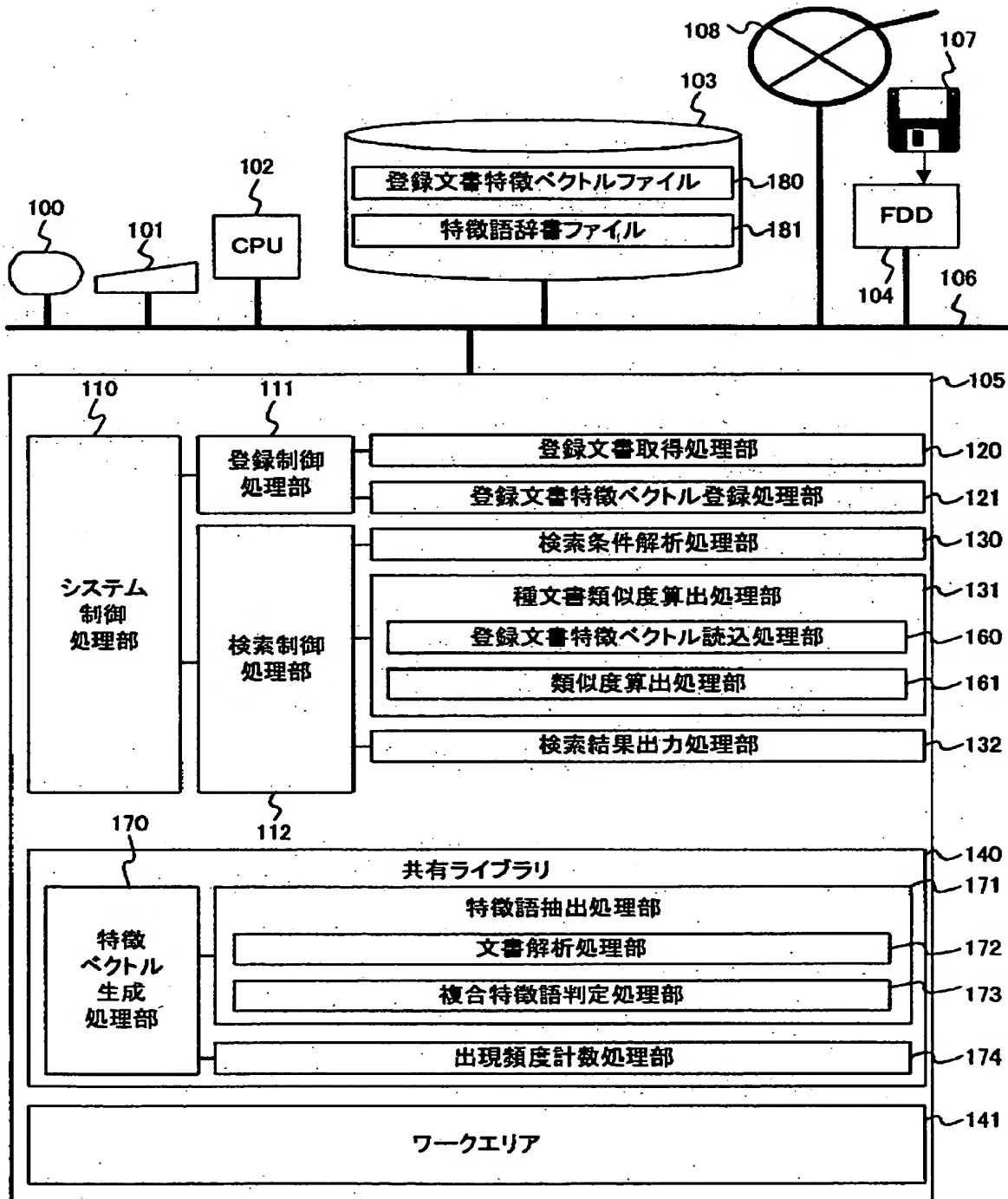
【符号の説明】

1 0 0 …ディスプレイ、1 0 1 …キーボード、1 0 2 …CPU、1 0 3 …磁気ディスク装置、1 0 4 …FDD、1 0 5 …主メモリ、1 0 6 …バス、1 0 7 …フロッピディスク、1 0 8 …ネットワーク、1 4 0 …共有ライブラリ、1 4 1 …ワークエリア、1 8 0 …登録文書特徴ベクトルファイル、1 8 1 …特徴語辞書ファイル、1 1 0 …システム制御処理部、1 1 1 …登録制御処理部、1 1 2 …検索制御処理部、1 2 0 …登録文書取得処理部、1 2 1 …登録文書特徴ベクトル登録処理部、1 3 0 …検索条件解析処理部、1 3 1 …種文書類似度算出処理部、1 3 2 …検索結果出力処理部、1 6 0 …登録文書特徴ベクトル読込処理部、1 6 1 …類似度算出処理部、1 7 0 …特徴ベクトル生成処理部、1 7 1 …特徴語抽出処理部、1 7 2 …文書解析処理部、1 7 3 …複合特徴語判定処理部、1 7 4 …出現頻度計数処理部、1 6 0 1 …種文書、1 6 0 2 …特徴語候補、1 6 0 2 及び 1 6 0 3 …登録文書特徴ベクトル、1 6 0 3 …特徴ベクトル、1 6 0 4 …特徴語、1 6 0 5 …特徴ベクトル、1 7 0 0 及び 1 7 0 1 …登録文書、1 7 0 2 及び 1 7 0 3 …登録文書特徴ベクトル、1 7 0 5 …種文書、1 7 0 6 …種文書特徴ベクトル、1 7 0 7 …類似度算出結果、1 9 0 0 …出現位置取得処理部、2 0 0 0 …重み係数算出処理部、2 3 0 0 及び 2 3 0 1 …登録文書特徴ベクトル、2 4 0 0 …種文書特徴ベクトル、2 4 0 1 …重み係数、2 4 0 2 …類似度算出結果、2 5 0 0 …特徴ベクトル、4 0 1 及び 4 0 2 …登録文書、4 0 3 及び 4 0 4 …登録文書特徴ベクトル、4 0 5 …特徴語辞書、4 0 6 …種文書、4 0 7 …種文書特徴ベクトル、4 0 8 …類似度算出結果。

【書類名】 図面

【図 1】

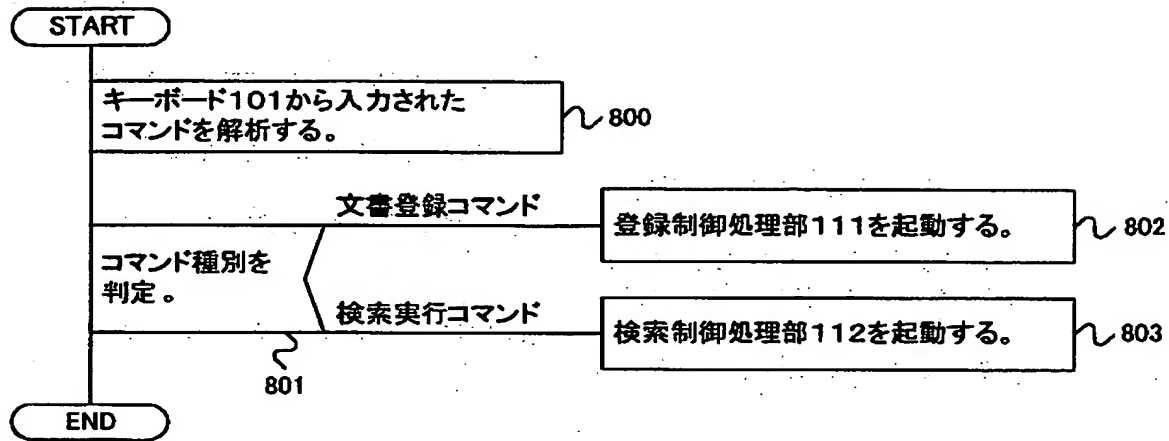
図 1





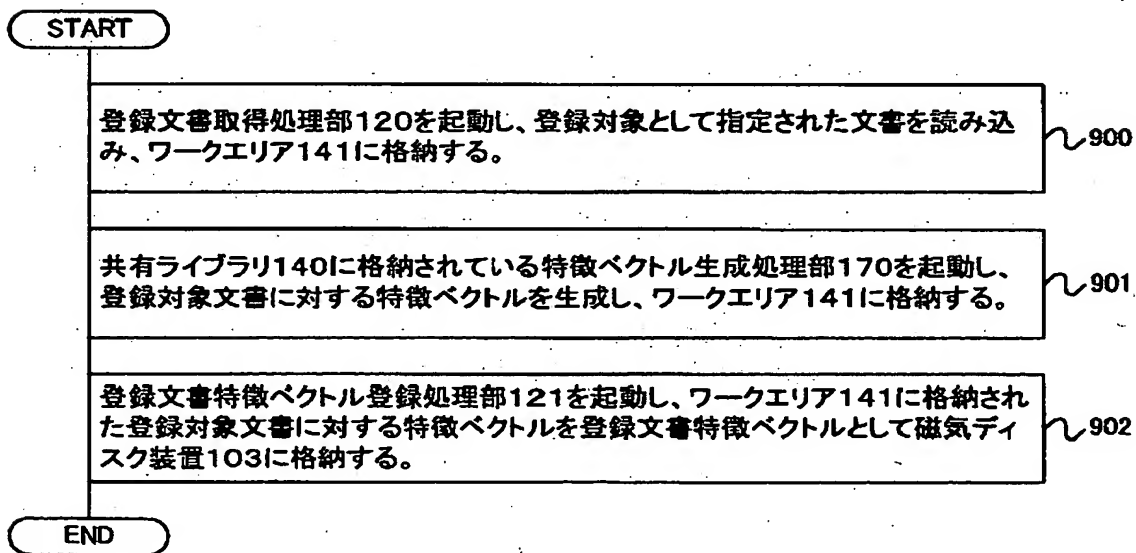
【図 2】

図2



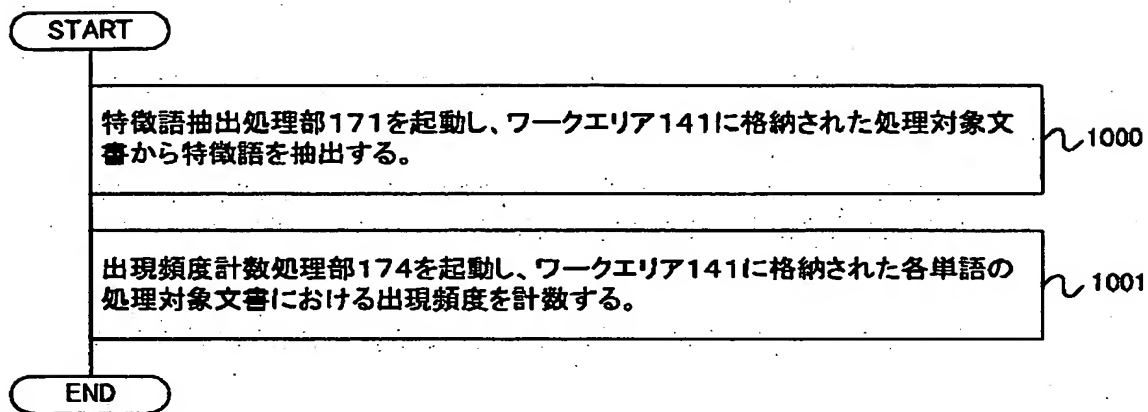
【図 3】

図3



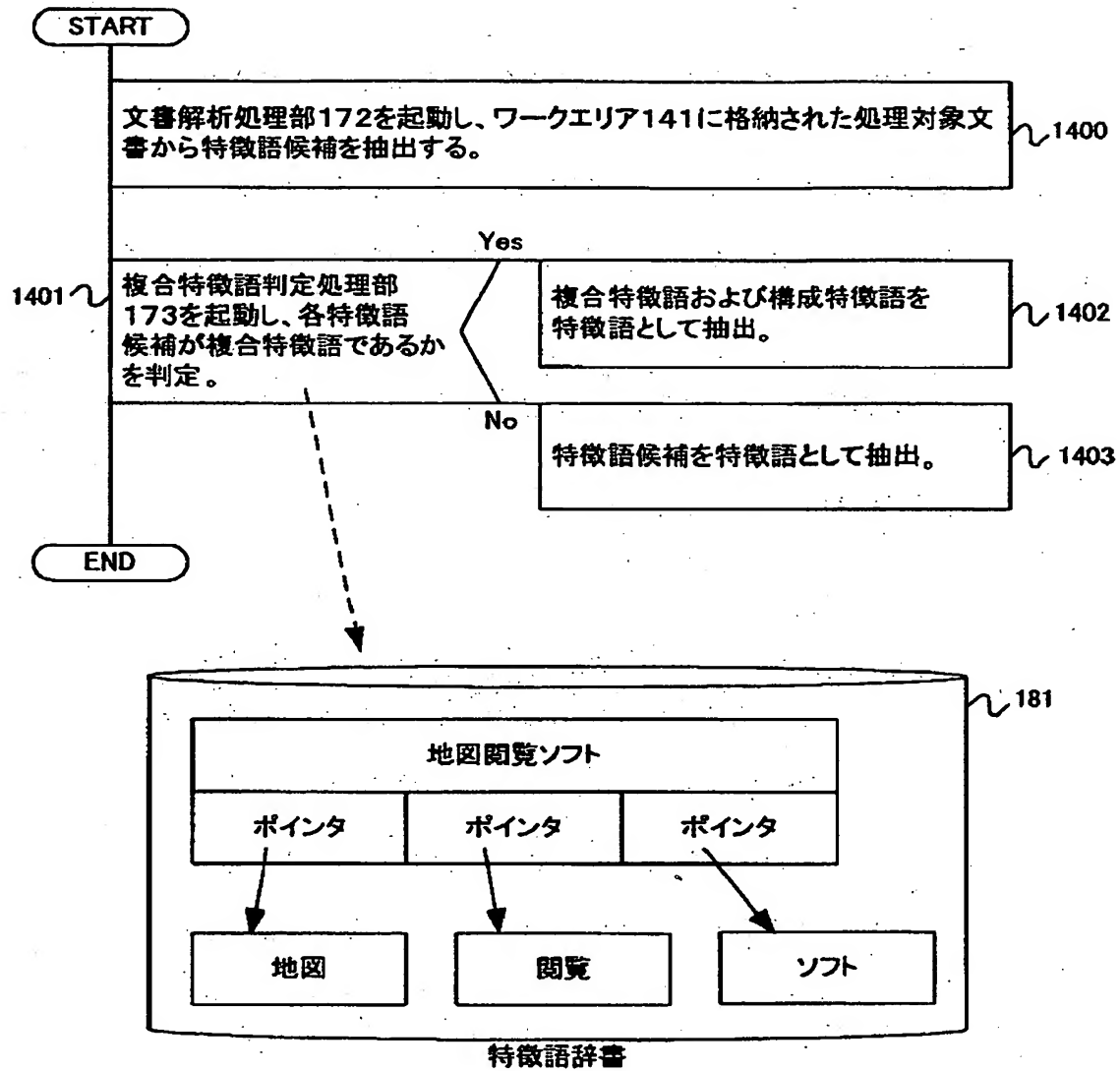
【図 4】

図 4



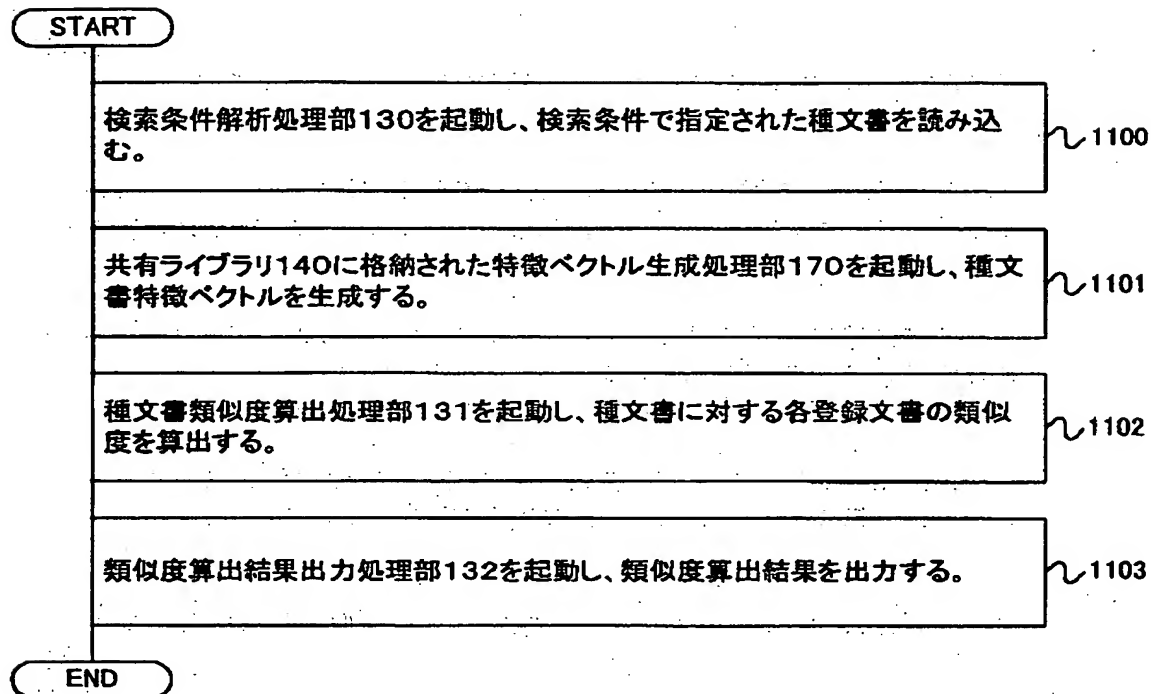
【図5】

図5



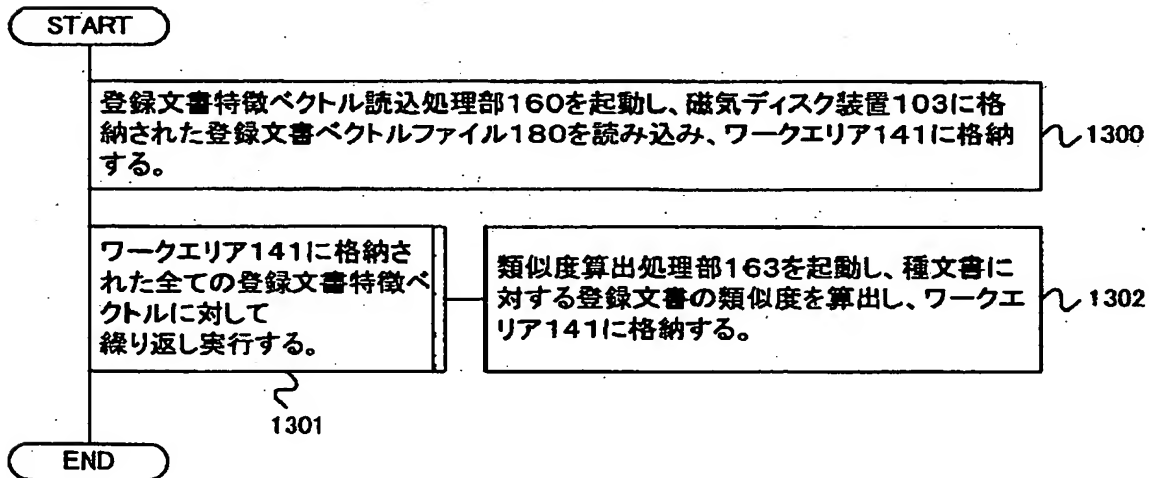
【図 6】

図6

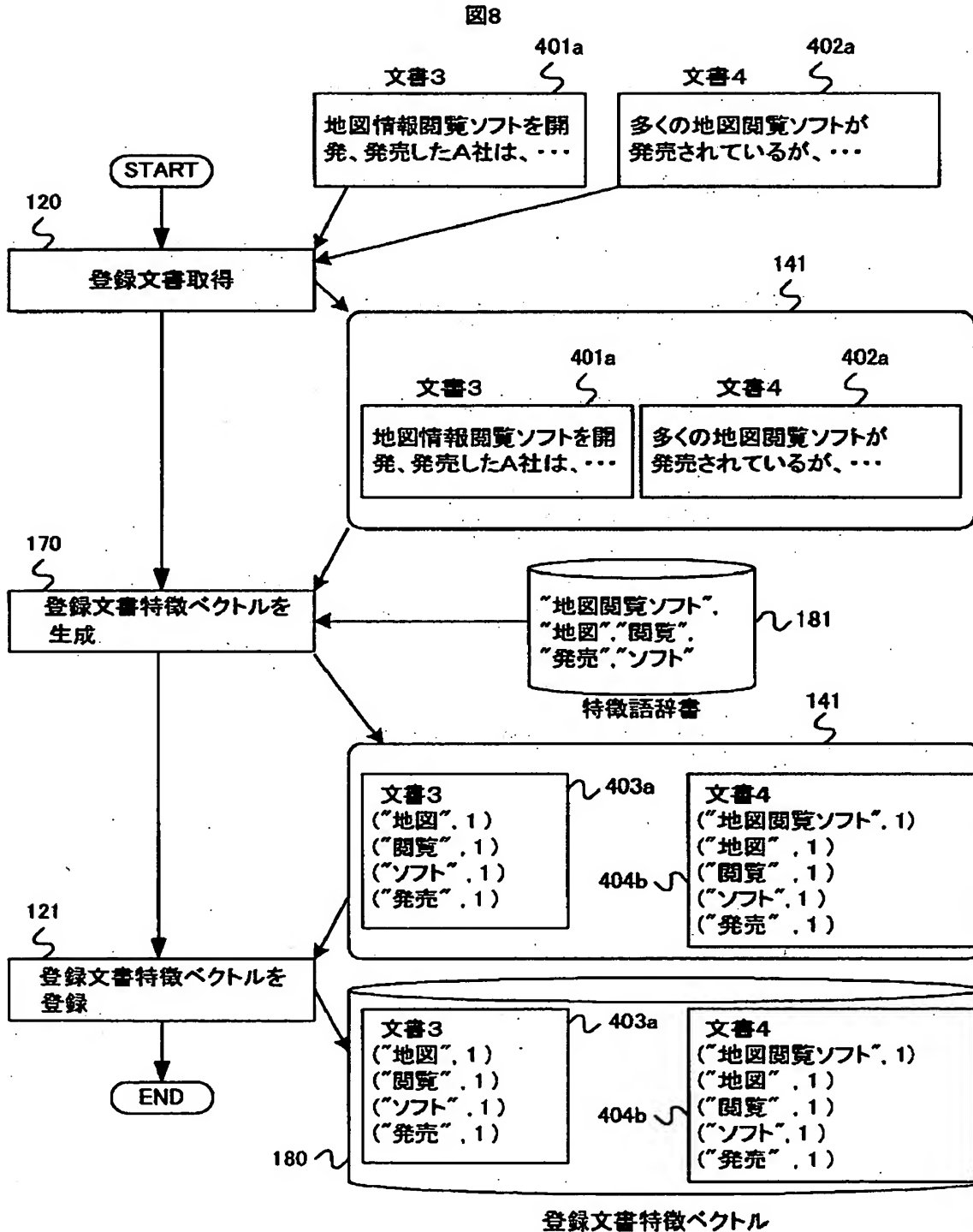


【図 7】

図7

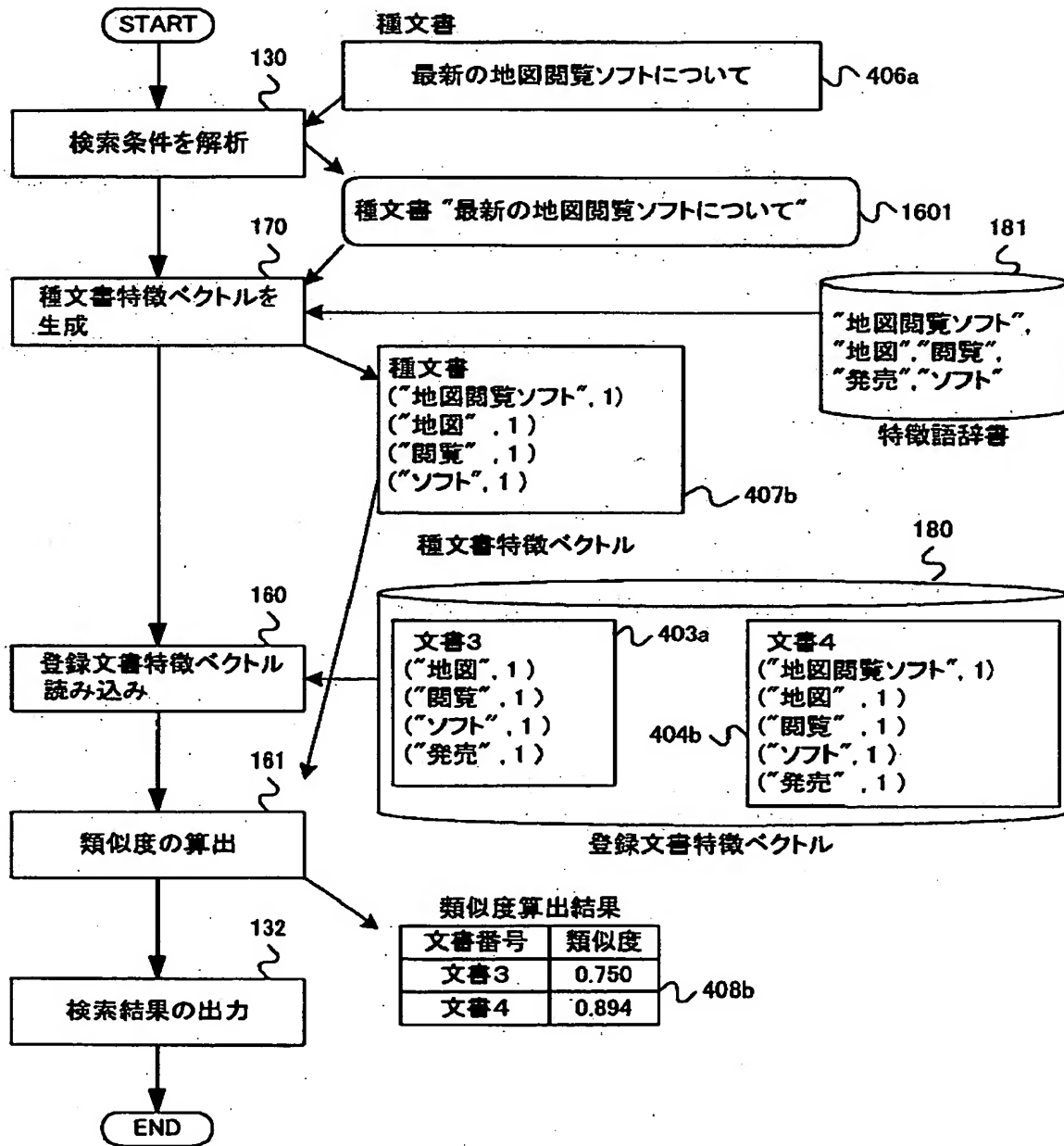


【図8】



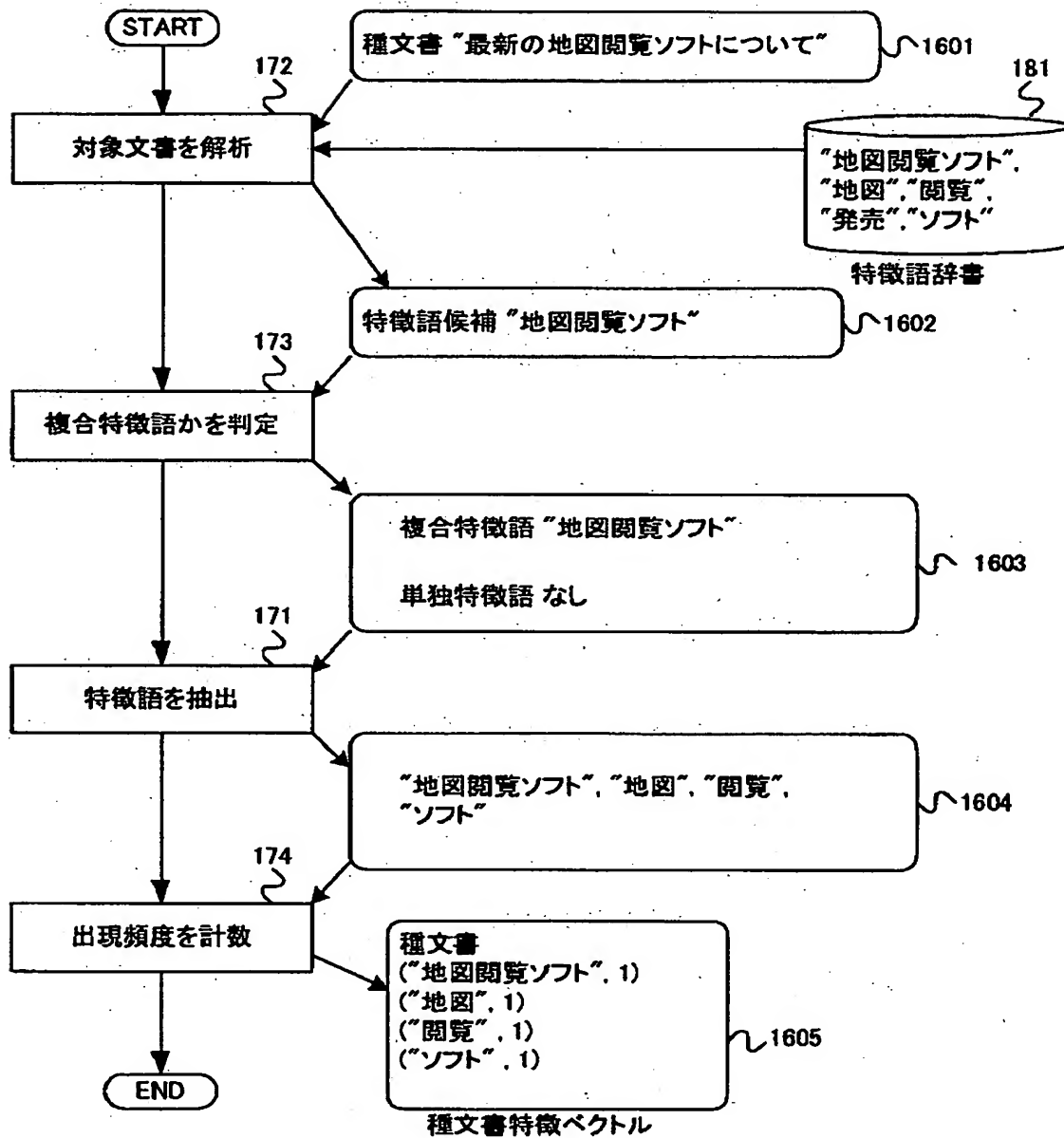
【図9】

図9



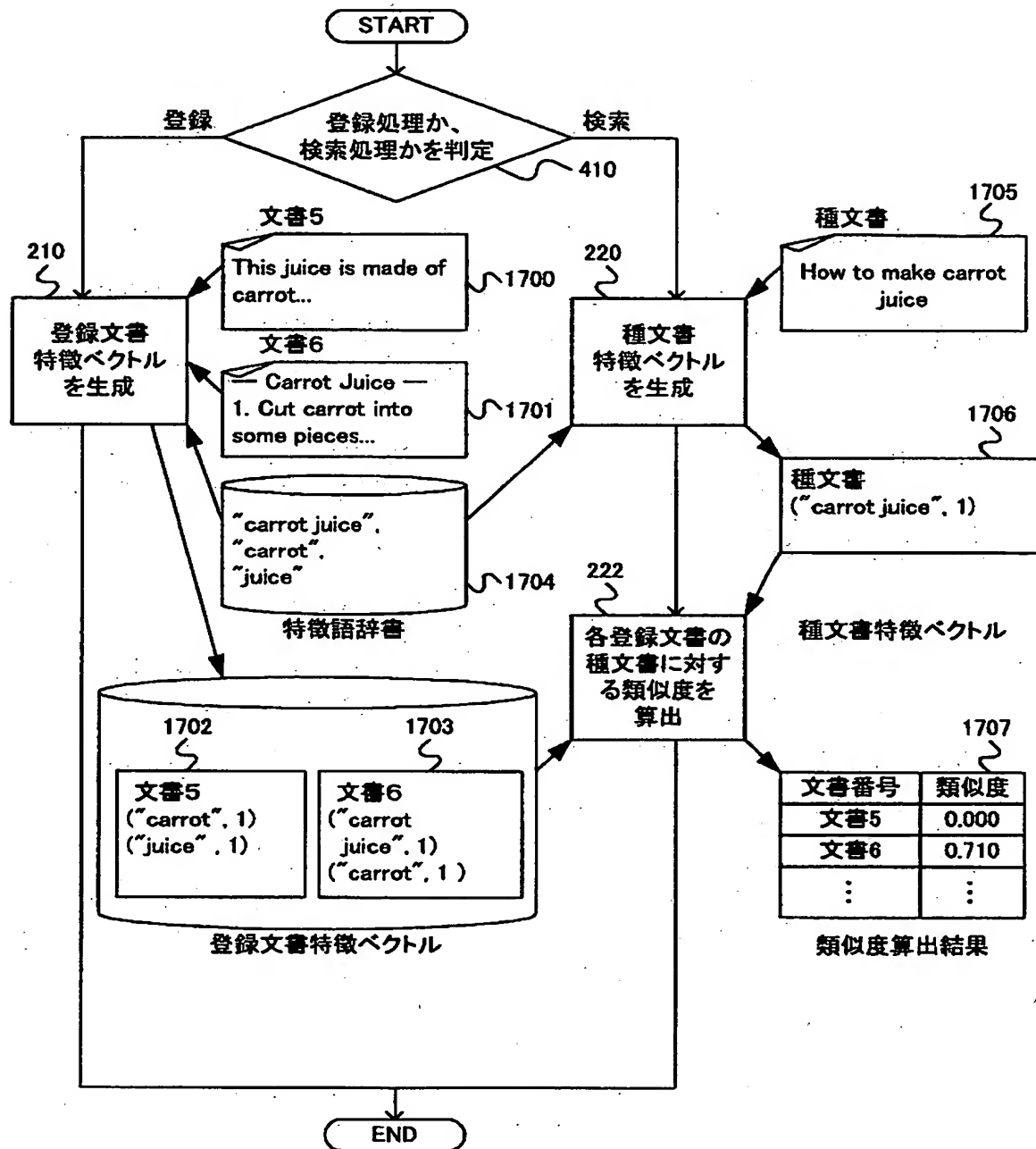
【図 1 0】

図10



【図 11】

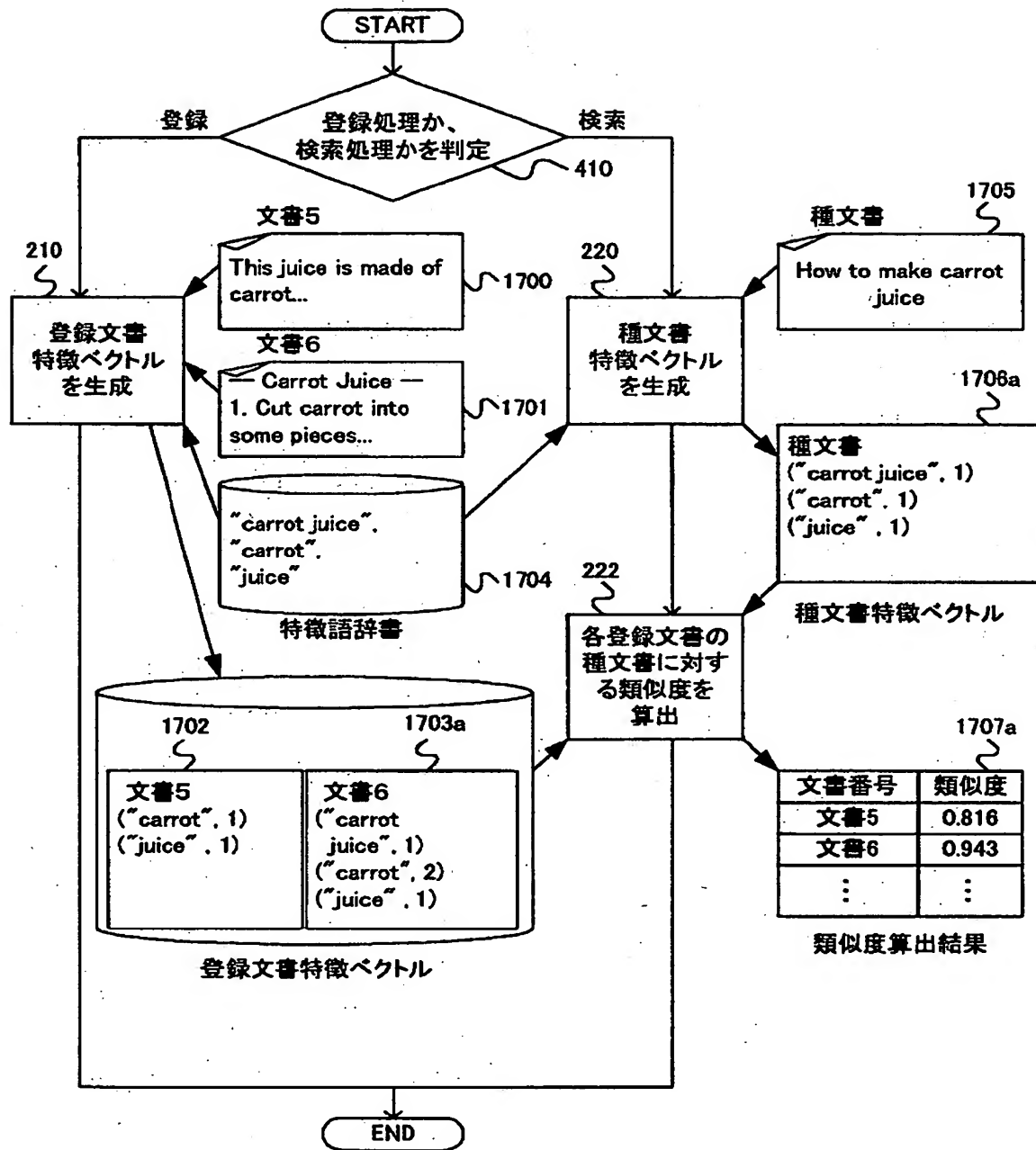
図11





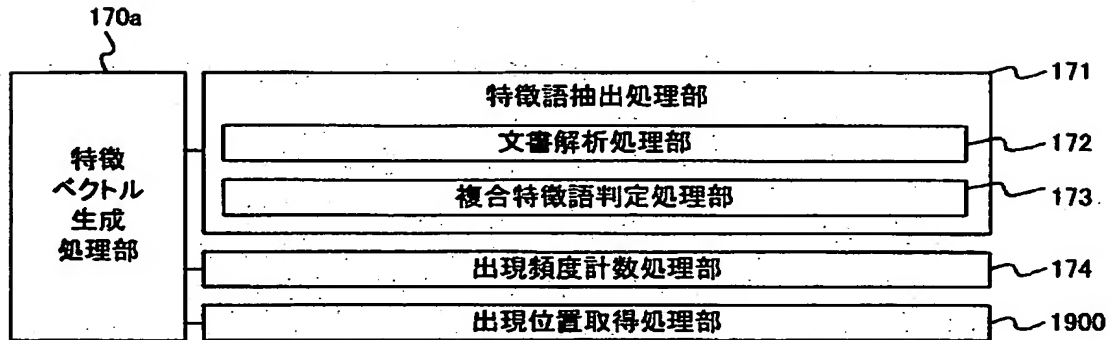
【図12】

図12



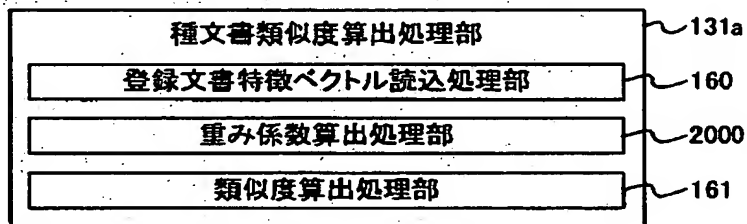
【図 1 3】

図13



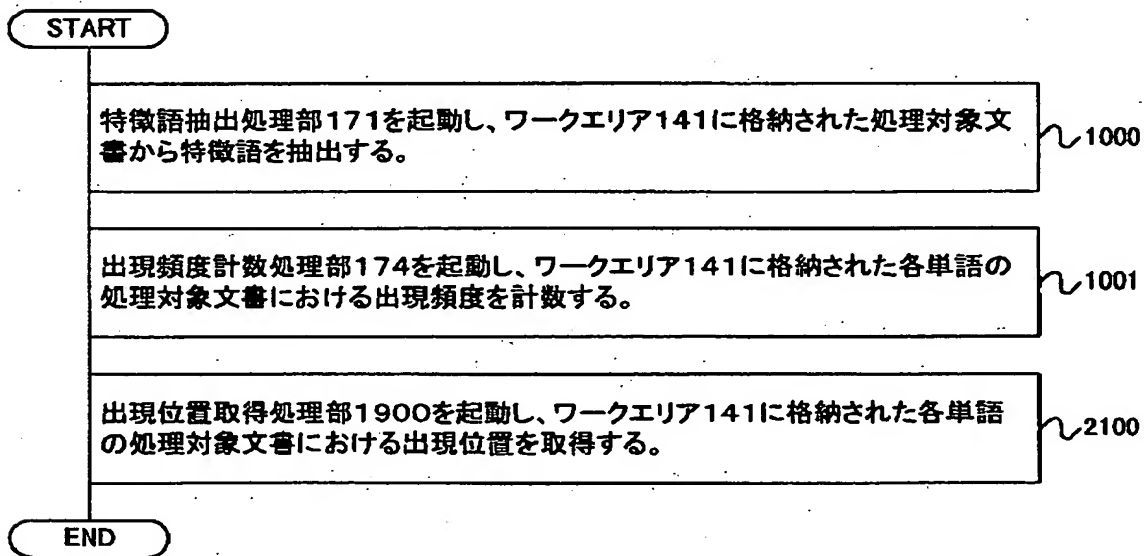
【図 1 4】

図14



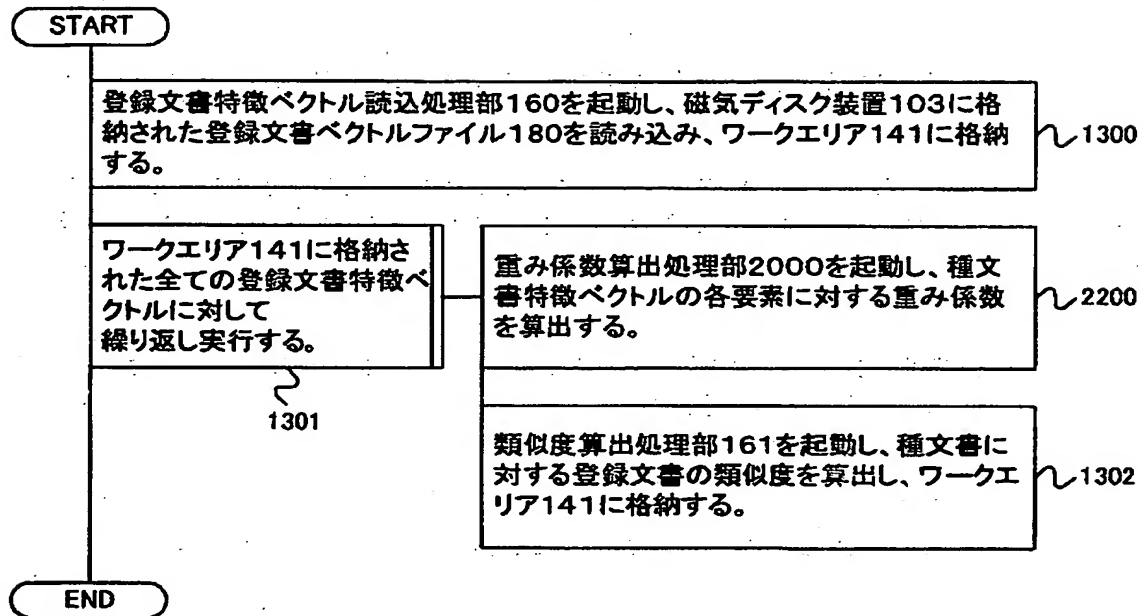
【図 1 5】

図15

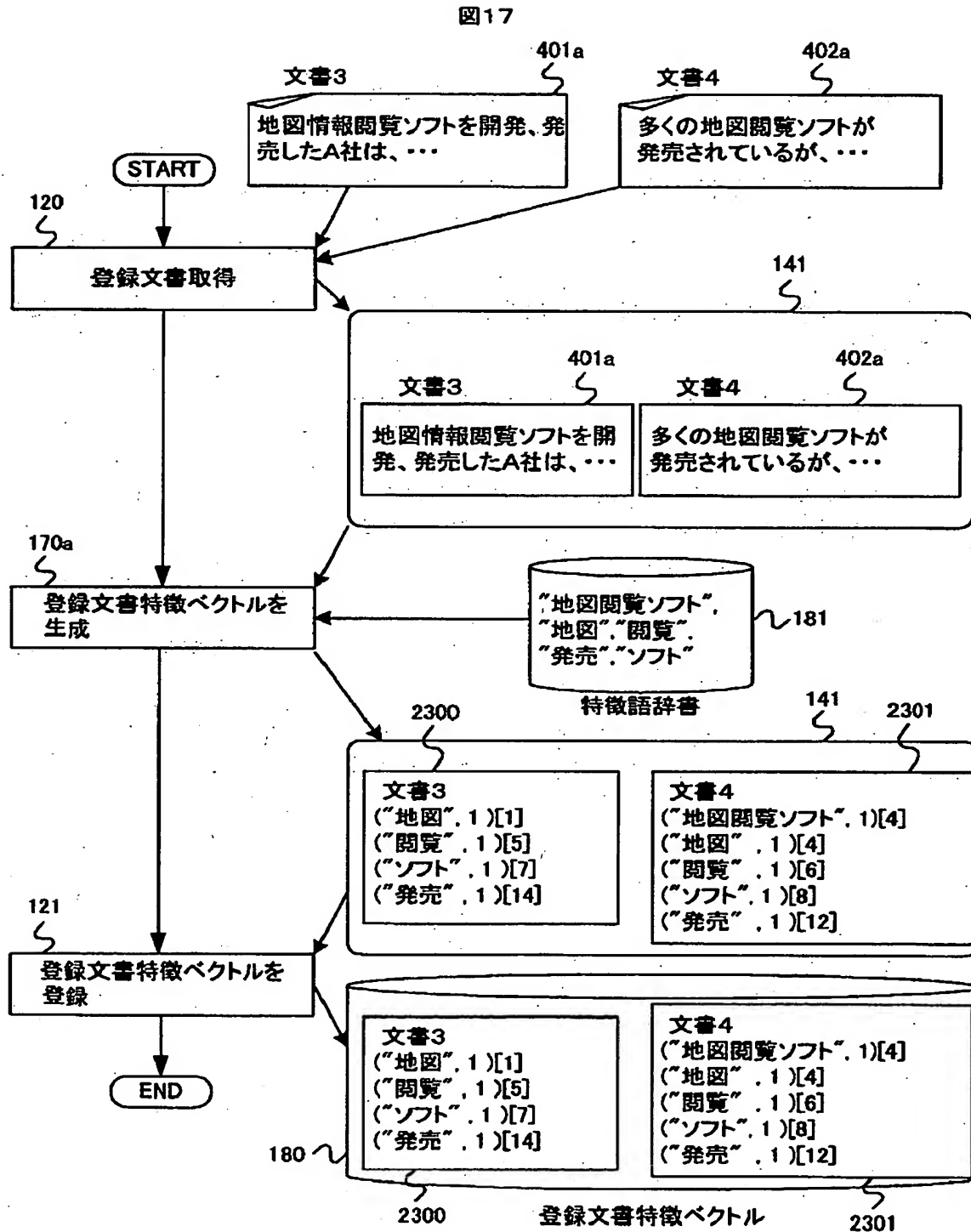


【図 1 6】

図16

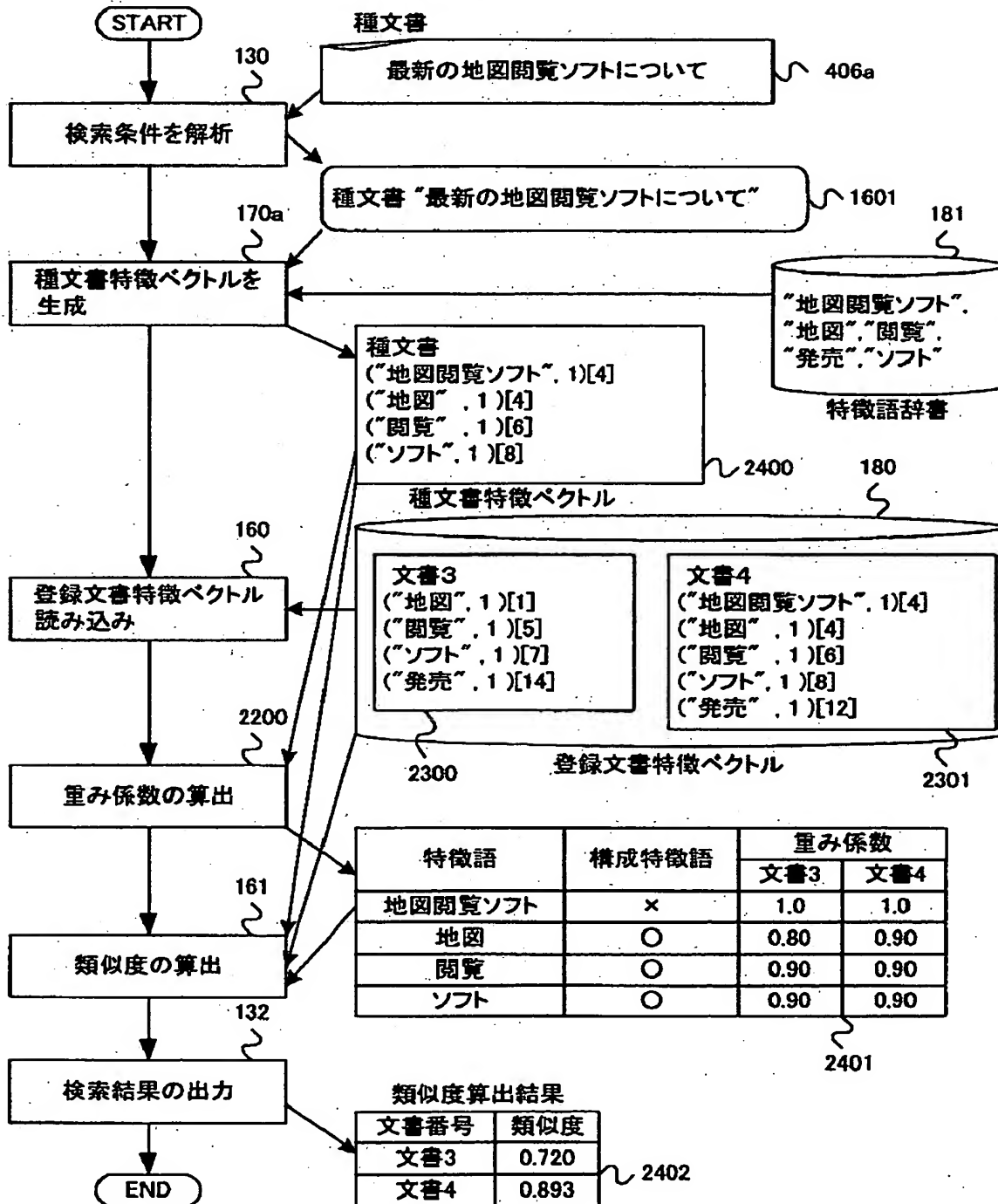


【図 17】



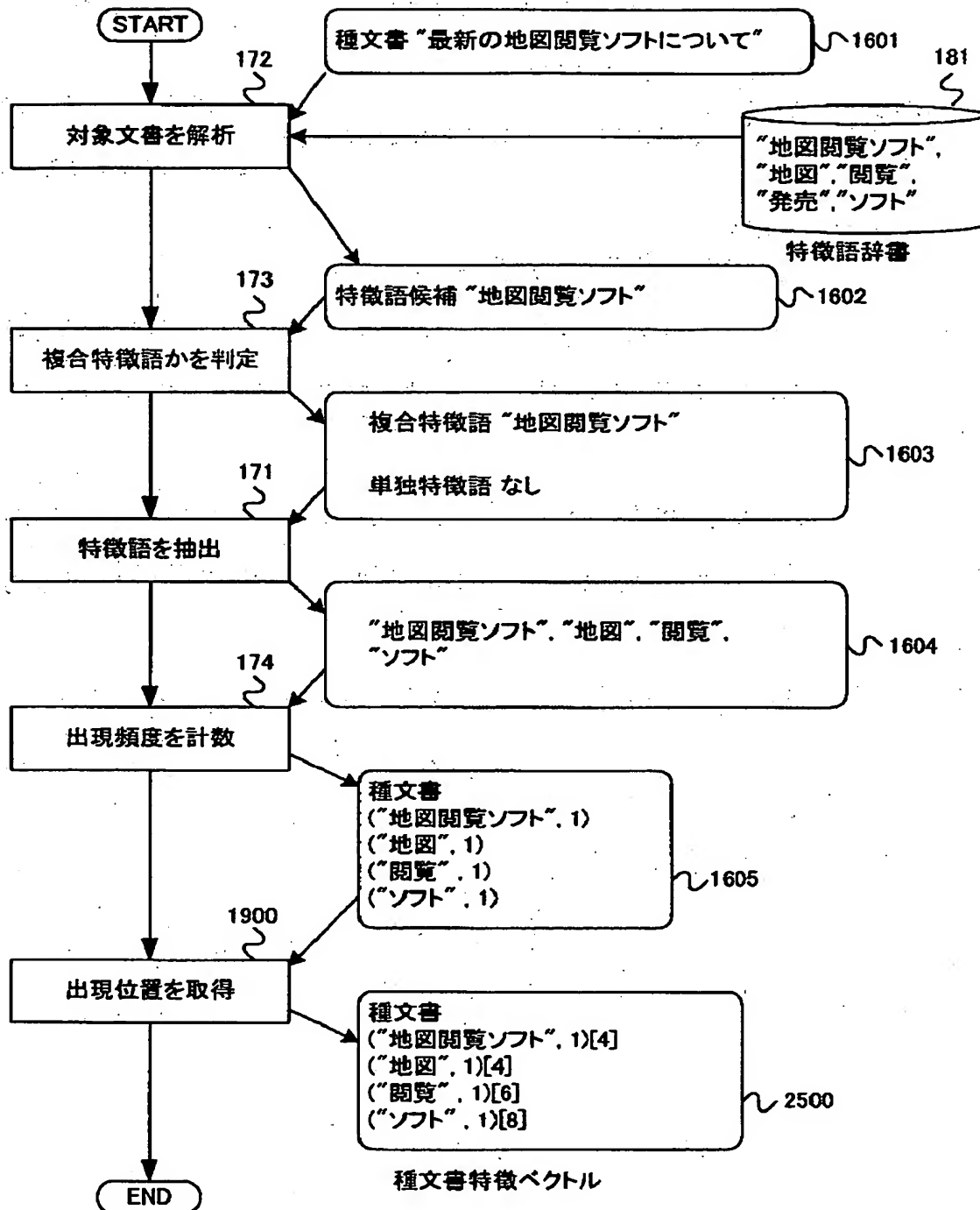
【図18】

図18

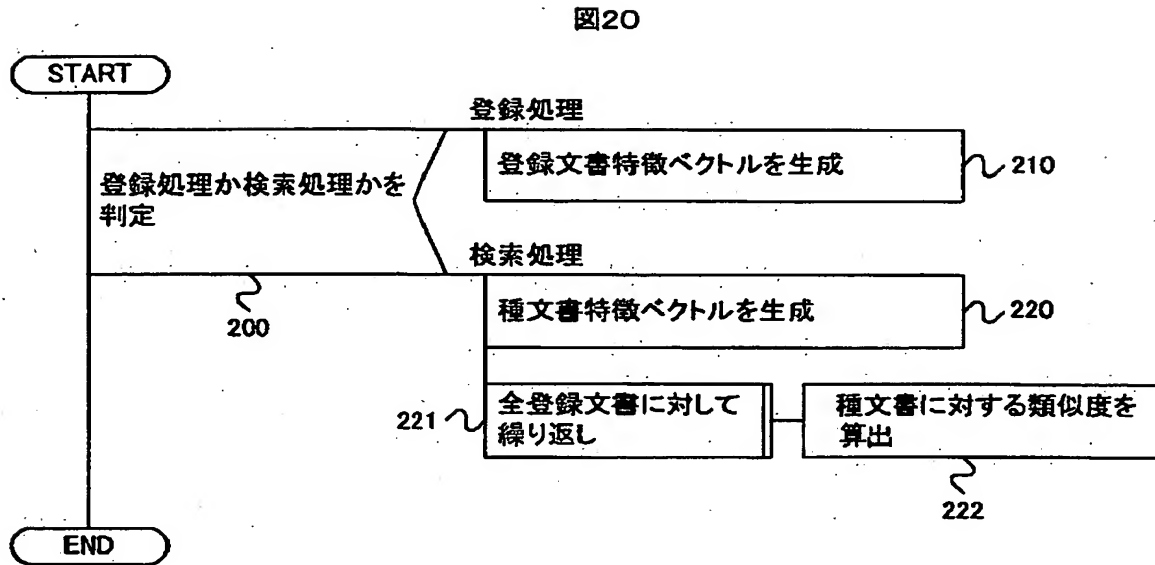


【図 19】

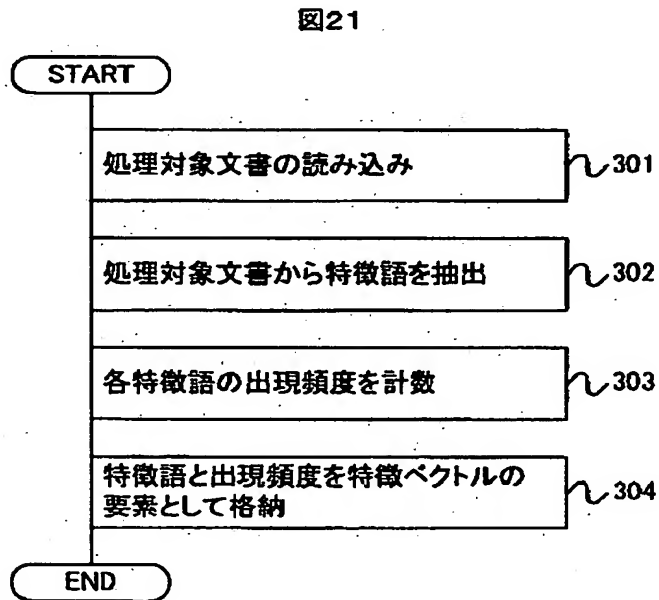
図 19



【図 2 0】

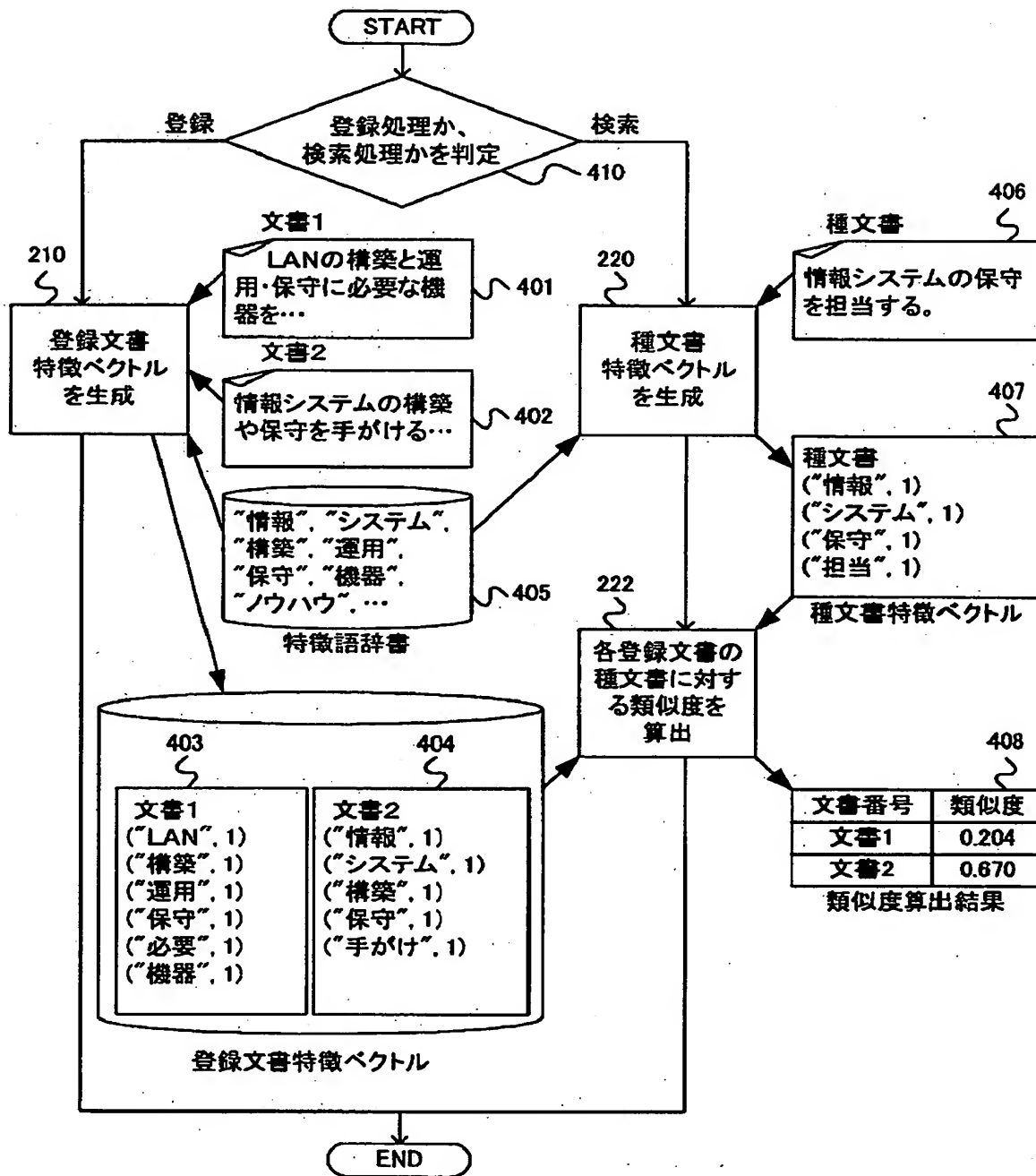


【図 2 1】



【図22】

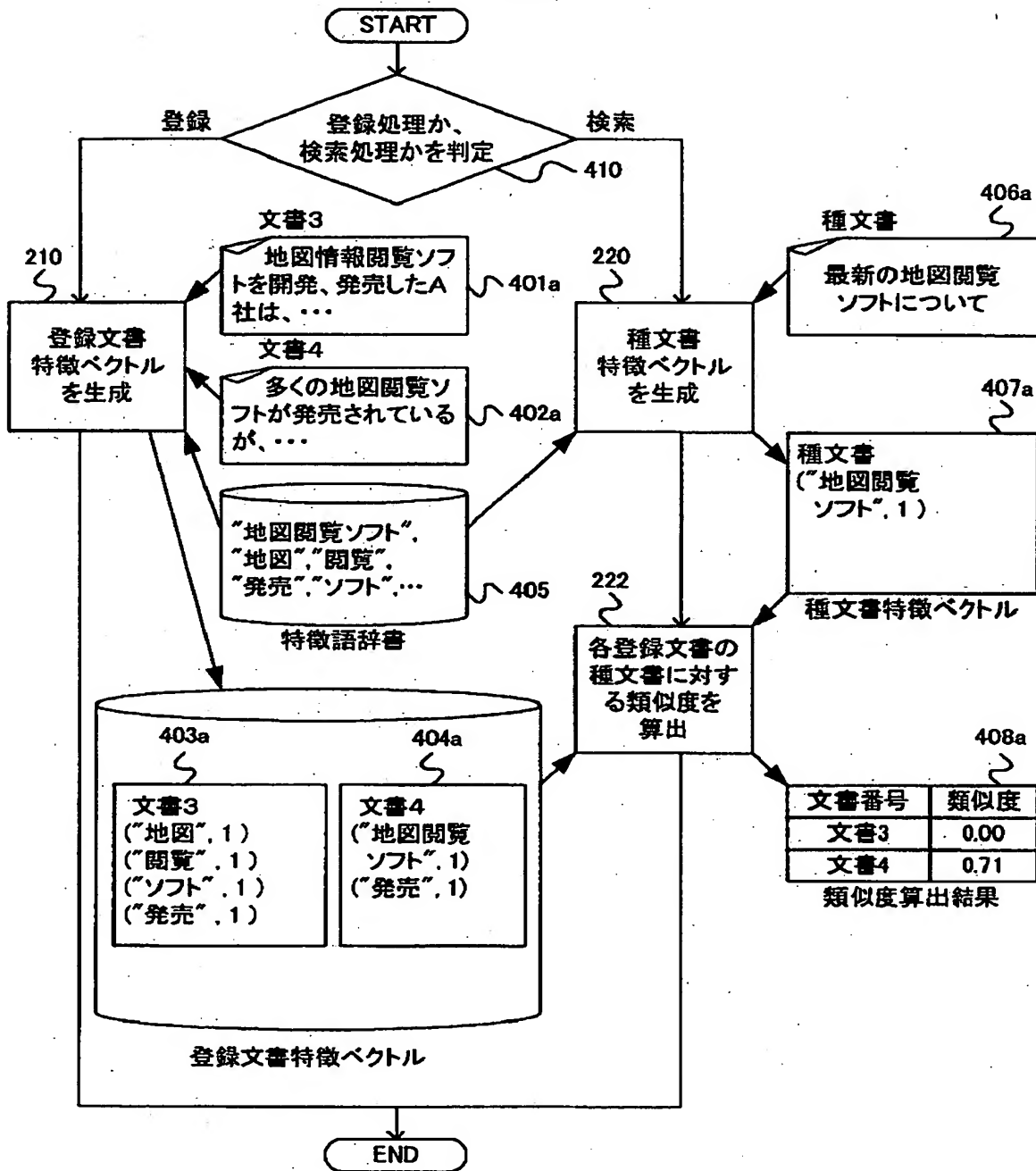
図22





【図23】

図23



【書類名】            要約書

【要約】

【課題】    検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を精度良く検索することが可能な技術を提供する。

【解決手段】    指定された文書と類似する文書を検索する類似文書検索方法において、所望の検索内容を含んだ種文書から特徴語の候補となる特徴語候補を抽出するステップと、前記抽出された特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該種文書の特徴語として抽出するステップと、前記抽出された種文書の特徴語と登録文書の特徴語との間の類似度を算出するステップと、前記算出された類似度算出結果を検索結果として出力するステップとを有するものである。

【選択図】            図 1

特2001-128934

## 認定・付加情報

特許出願の番号	特願2001-128934
受付番号	50100616231
書類名	特許願
担当官	第七担当上席 0096
作成日	平成13年 5月 2日

### <認定情報・付加情報>

【提出日】 平成13年 4月26日

次頁無

出願人履歴情報

識別番号 [000005108]

1. 変更年月日 1990年 8月31日

[変更理由] 新規登録

住 所 東京都千代田区神田駿河台4丁目6番地  
氏 名 株式会社日立製作所